

Paweł Lula

Akademia Ekonomiczna w Krakowie

PODOBIEŃSTWO OBIEKTÓW OPISYWANYCH ZA POMOCĄ STRUKTUR DRZEWIASTYCH

1. Wstęp

Jednym z najszerzej wykorzystywanych obecnie narzędzi pozwalających na przechowywanie, przetwarzanie i przesyłanie danych w systemach informatycznych jest język XML. Jego zastosowanie pozwala na:

- definiowanie struktury danych i badanie zgodności przechowywanych danych z określoną wcześniej strukturą,
- przechowywanie danych w systemach komputerowych,
- wymianę danych pomiędzy różnymi systemami,
- implementowanie algorytmów przetwarzania zgromadzonych zasobów,
- konwersję zgromadzonych zasobów do różnorodnych postaci (np. HTML, PDF, format tekstowy oraz wiele innych).

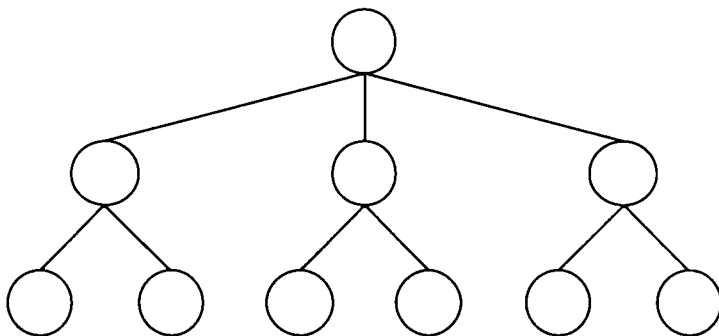
Podstawową złożoną strukturą danych stosowanych w języku XML jest drzewo. Drzewem jest cały dokument XML, natomiast poszczególne obiekty opisane w dokumencie przyjmują formę poddrzew będących gałęziami drzewa dokumentu. Każde poddrzewo może zawierać kolejne struktury drzewiaste jako swoje elementy składowe. W szczególnym przypadku wszystkie elementy potomne drzewa mogą być liśćmi, które nie mają już dalszych potomków – wówczas mogą być one rozpatrywane jako listy. Wysoki stopień upowszechnienia struktur drzewiastych w systemach informacyjnych stwarza konieczność analizy przechowywanych w nich informacji. Jednym z najczęściej spotykanych typów analizy danych są badania porównawcze, których najważniejszym elementem jest określenie podobieństwa pomiędzy rozpatrywanymi obiektami. W przypadku przechowywania informacji w formacie XML zadanie to sprowadza się do wyznaczenia podobieństwa pomiędzy drzewami reprezentującymi rozpatrywane obiekty.

Zasadniczym celem niniejszej pracy jest zaprezentowanie propozycji metody wyznaczania podobieństwa pomiędzy drzewami zapisanymi w dokumencie XML. Potrzeba prowadzenia badań w tym zakresie wynika z następujących przesłanek:

- język XML jest stosowany na szeroką skalę do przechowywania danych o złożonej strukturze; próba automatyzacji procesu ich analizy wymaga zdefiniowania sposobu ich przetwarzania w trakcie realizacji obliczeń, a problem wyznaczania podobieństwa należy do najistotniejszych, rozpatrywanych w tym kontekście, zagadnień badawczych;
- język XML jest wykorzystywany na coraz szerszą skalę przy tworzeniu serwisów internetowych. Z tego powodu opracowanie metod analizy danych reprezentowanych za pomocą struktur drzewiastych jest elementem niezbędnym przy podejmowanych próbach pozyskiwania informacji z witryn WWW i dalszego jej przetwarzania;
- XML staje się podstawowym formatem przechowywania danych dla wielu typów aplikacji, w tym również aplikacji biurowych (edytory, arkusze kalkulacyjne, programy prezentacyjne). Również w tym przypadku analiza tego typu informacji wymaga opracowania zasad ich uwzględniania w obliczeniach;
- XML jest stosowany w systemach elektronicznej wymiany dokumentów, które również mogą być poddawane automatycznej analizie, w tym określaniu podobieństwa pomiędzy całymi dokumentami lub pomiędzy ich fragmentami.

2. Struktury drzewiaste i ich reprezentacja w dokumencie XML

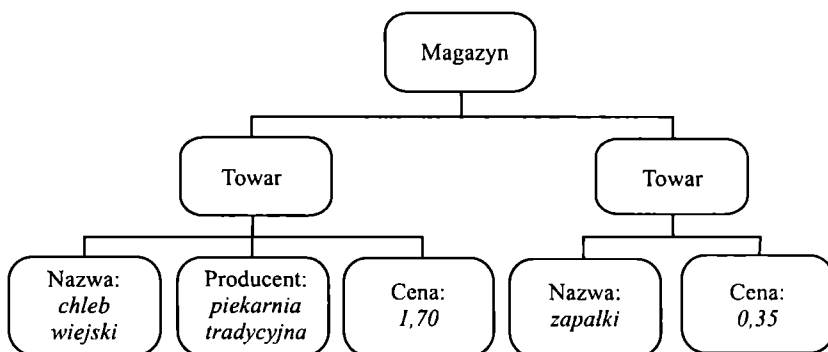
Drzewo jest jedną z podstawowych struktur złożonych. Jest ono definiowane jako spójny, acykliczny graf nieskierowany [Cormen i in. 2004]. Inna definicja drzewa mówi, że drzewo to kolekcja elementów pozostających w zależności hierarchicznej posiadająca jeden element wyróżniony, zwany korzeniem drzewa.



Rys. 1. Przykładowe drzewo

Źródło: opracowanie własne.

Drzewo jest dogodnym narzędziem opisu obiektów o strukturze złożonej, powszechnie wykorzystywanym w systemach informatycznych. W roli ilustracji przedstawione zostanie drzewo prezentujące prosty magazyn towarów wraz z zawartością.



Rys. 2. Opis magazynu wraz z zawartością za pomocą drzewa

Źródło: opracowanie własne.

Definicja drzewa reprezentującego przykładowy magazyn w języku XML przyjmuje postać:

```

<!DOCTYPE MAGAZYN [
  <!ELEMENT NAZWA (#PCDATA)>
  <!ELEMENT PRODUCENT (#PCDATA)>
  <!ELEMENT CENA (#PCDATA)>
  <!ELEMENT TOWAR (NAZWA, PRODUCENT?, CENA)>
  <!ELEMENT MAGAZYN (TOWAR+)>
]>,
  
```

zawartość magazynu zaś może zostać zapisana w dokumencie XML w następującej formie:

```

<MAGAZYN>
  <TOWAR>
    <NAZWA>Chleb wiejski</NAZWA>
    <PRODUCENT>Piekarnia Tradycyjna</PRODUCENT>
    <CENA>1.70</CENA>
  </TOWAR>
  <TOWAR>
    <NAZWA>Zapalki</NAZWA>
    <CENA>0.35</CENA>
  </TOWAR>
</MAGAZYN>.
  
```

Zamieszczony powyżej opis magazynu wykorzystany zostanie w dalszej części artykułu do ilustracji prezentowanych problemów dotyczących wyznaczania miary podobieństwa obiektów drzewiastych.

3. Podobieństwo drzew XML

Ze względu na duże upowszechnienie drzew w systemach gromadzenia i przetwarzania informacji rozważania dotyczące wyznaczania podobieństwa tego typu obiektów pojawiają się dość często w literaturze. Początkowe prace dotyczyły wyłącznie oceny podobieństwa strukturalnego (*structural similarity*) bazującego na porównaniu wyłącznie struktury dokumentu, a nie jego zawartości. Tego typu badania opierały się na analizie zapisów w plikach zawierających definicję typu dokumentu. Rozważania dotyczące podobieństwa rozumianego w sposób kompleksowy pojawiły się znacznie później. Jedną z najciekawszych prac z tego zakresu jest [Bhavsar i in. 2004].

Bezpośrednią inspiracją do rozważań zamieszczonych w niniejszej publikacji jest praca [Long i in. 2005] dotycząca wyznaczania odległości pomiędzy dokumentami XML.

Przyjęto, że propozycja prezentowanego miernika podobieństwa powinna spełniać następujące wymagania:

- miernik powinien uwzględniać różnice w strukturze obiektów oraz zróżnicowanie poszczególnych wartości,
- proponowany wskaźnik powinien uwzględniać wartości wyrażone na wszystkich czterech skalach pomiarowych oraz wartości tekstowe (miernik zaprezentowany w [Long i in. 2005] jest przystosowany wyłącznie do analizy wartości numerycznych i nominalnych);
- oczekuje się, że miernik przyjmować będzie wartość z przedziału $[0, 1]$, przy czym wartość jednostkowa oznacza pełne podobieństwo, a wartość zerowa świadczy o braku podobieństwa.

Przy wyznaczaniu miernika niezbędne jest opracowanie metod pomiaru odległości pomiędzy elementami zarówno o strukturze prostej (czyli takimi, które przechowują pojedynczą wartość), jak i o strukturze złożonej. Zagadnienia te zostaną przedstawione w dwóch kolejnych punktach pracy.

4. Podobieństwo elementów prostych

Przez element prosty rozumie się parę $\langle n_i, w_i \rangle$, gdzie n_i jest nazwą elementu, w_i zaś jego wartością. Przyjmuje się, że w_i jest wartością wyrażoną na jednej z czterech podstawowych skal pomiarowych lub też jest wartością o charakterze tekstowym. W szczególnym przypadku w_i może być wartością pustą (w dalszej części rozważań wartości puste oznaczone będą przez *EMPTY*).

Proponuje się przyjęcie następujących reguł wyznaczania podobieństwa pomiędzy elementami prostymi:

a) podobieństwo elementów prostych o różnych nazwach

Podobieństwo pomiędzy elementami o różnych nazwach jest na poziomie zerowym, co można zapisać jako:

$$p_{XML}(\langle n_i, w_i \rangle, \langle n_j, w_j \rangle) = 0, \text{ gdy } n_i \neq n_j. \quad (1)$$

Odwołując się do zaprezentowanego wcześniej dokumentu, można stwierdzić, że:

$$p_{XML}(\langle \text{NAZWA} \rangle \text{Zapałki} \langle / \text{NAZWA} \rangle, \langle \text{CENA} \rangle 12,50 \langle / \text{CENA} \rangle) = 0.$$

b) podobieństwo elementów o wartościach pustych (EMPTY)

Analizując dwa elementy o identycznych nazwach, w których w obu przypadkach wartości są elementami pustymi, można stwierdzić, że ich podobieństwo jest całkowite:

$$p_{XML}(\langle n_i, \text{EMPTY} \rangle, \langle n_i, \text{EMPTY} \rangle) = 1. \quad (2)$$

Natomiast gdy wartość pusta występuje tylko w jednym elemencie, to podobieństwo pomiędzy elementami nie występuje:

$$p_{XML}(\langle n_i, w_i \rangle, \langle n_i, \text{EMPTY} \rangle) = p_{XML}(\langle n_i, \text{EMPTY} \rangle, \langle n_i, w_i \rangle) = 0. \quad (3)$$

Za przykład takiej sytuacji służyć może zapis:

$$p_{XML}(\langle \text{NAZWA} \rangle \text{Zapałki} \langle / \text{NAZWA} \rangle, \text{EMPTY}) = 0.$$

Kolejne przypadki dotyczyć będą sytuacji, w których w obu porównywanych elementach występować będą identyczne nazwy i niepuste wartości.

c) podobieństwo elementów o wartościach wyrażonych na skali nominalnej

Jeśli wartości przechowywane w elementach wyrażone są na skali nominalnej, to miernik podobieństwa pomiędzy elementami przyjmuje jedną z dwóch wartości.

$$p_{XML} = \begin{cases} 1 & \text{gdy } w_i = w_j \\ 0 & \text{gdy } w_i \neq w_j \end{cases}. \quad (4)$$

Korzystając z powyższej formuły, można stwierdzić, że:

$$p_{XML}(\langle \text{NAZWA} \rangle \text{Zapałki} \langle / \text{NAZWA} \rangle, \langle \text{NAZWA} \rangle \text{Chleb} \langle / \text{NAZWA} \rangle) = 0.$$

d) podobieństwo elementów o wartościach wyrażonych na skali porządkowej

Jeżeli wartości porównywanych elementów wyrażone są na skali porządkowej i przyjmują kolejne wartości $w_0, w_1, \dots, w_n$, to do określenia wartości miernika podobieństwa można wykorzystać wzór:

$$p_{XML} = 1 - \frac{|i - j|}{n}. \quad (5)$$

Na przykład jeśli wiadomo, że ocena może przyjąć wartości: *bardzo dobry*, *dobry*, *dostateczny*, *niedostateczny*, to:

$$p_{XML}(\langle \text{OCENA} \rangle \text{bardzo dobry} \langle / \text{OCENA} \rangle, \langle \text{OCENA} \rangle \text{dobry} \langle / \text{OCENA} \rangle) = 2/3.$$

e) podobieństwo elementów o wartościach wyrażonych na skali przedziałowej

Chcąc obliczyć wartość miernika podobieństwa pomiędzy elementami o wartościach przyjmujących wartości wyrażone na skali przedziałowej i mieszczące się w przedziale od w_{min} do w_{max} , można zastosować formułę:

$$p_{XML} = 1 - \frac{|w_i - w_j|}{w_{max} - w_{min}}. \quad (6)$$

Przyjmijmy np., że element *temperatura* może przyjmować wartości z przedziału $\langle -30; 50 \rangle$. Wówczas:

$$p_{XML}(\langle \text{TEMPERATURA} \rangle 20 \langle / \text{TEMPERATURA} \rangle, \langle \text{TEMPERATURA} \rangle -10 \langle / \text{TEMPERATURA} \rangle) = 5/8.$$

f) podobieństwo elementów o wartościach wyrażonych na skali ilorazowej

W tym przypadku dla rozpatrywanej wartości należy znać wartość maksymalną (wartość minimalna jest zawsze równa zero). Podobieństwo dane jest formułą:

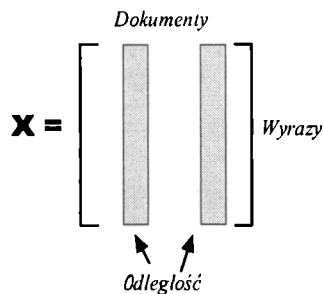
$$p_{XML} = 1 - \frac{|w_i - w_j|}{w_{max}}. \quad (7)$$

Niech przykładowa formuła dotyczy ceny towarów (przyjmującej wartość maksymalną równą 80):

$$p_{XML}(\langle \text{CENA} \rangle 20 \langle / \text{CENA} \rangle, \langle \text{CENA} \rangle 10 \langle / \text{CENA} \rangle) = 7/8.$$

g) podobieństwo elementów o wartościach w postaci tekstowej

W przypadku elementów przechowujących dane tekstowe niezbędne jest wyznaczenie miary podobieństwa tekstów. Sposób jej wyznaczania jest uzależniony od przyjętego sposobu reprezentacji tekstów. Najpopularniejsza metoda reprezentacji zakłada, że dokument jest reprezentowany przez wektor przechowujący informacje o częstości występowania zawartych w nim wyrazów. W przypadku analizy większej liczby dokumentów do ich reprezentacji wykorzystywana jest *macierz częstości* o postaci przedstawionej na rys. 3.



Rys. 3. Schemat reprezentacji dokumentów w postaci macierzy częstości

Elementy x_{ij} określają, ile razy i -ty wyraz występuje w j -tym dokumencie. Kolejne kolumny macierzy reprezentują rozpatrywane dokumenty, wiersze zaś odnoszą się do kolejnych wyrazów występujących w przetwarzanym zbiorze dokumentów. Takie podejście pozwala na zdefiniowanie miernika odległości pomiędzy dokumentami w postaci odległości reprezentujących je wektorów. W analogiczny sposób rozpatrywać można odległość pomiędzy wyrazami. Będzie nią odległość pomiędzy wierszami macierzy częstości odpowiadającymi porównywanym wyrazom. W stosunku do przedstawionej metody reprezentacji można sformułować wiele zarzutów. Autorzy najpoważniejszego z nich stwierdzają, że stosowanie macierzy częstości do reprezentacji dokumentów prowadzi do pominięcia informacji dotyczącej kolejności występowania wyrazów w dokumencie. Fakt ten jest istotną ułomnością tej metody reprezentacji. Kolejną uwagą dotyczącą stosowania macierzy częstości jest stwierdzenie głoszące, że na wyniki przeprowadzanych analiz większy wpływ może mieć informacja o dokładnej liczbie wystąpień wyrazu niż samo jego wystąpienie w tekście. Często spotykaną metodą rozwiązania tego problemu jest przekształcenie macierzy częstości do macierzy binarnej poprzez zastąpienie elementów niezerowych jedynekami.

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & 4 & \dots & 4 \\ 1 & 0 & 3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 2 & \dots & 1 \end{bmatrix} \Rightarrow \mathbf{X}^{bin} = \begin{bmatrix} 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 1 & \dots & 1 \end{bmatrix}$$

Rys. 4. Schemat tworzenia macierzy binarnej

Źródło: opracowanie własne.

Zastosowanie macierzy binarnej jest jedną z wielu metod przekształcenia macierzy częstości. Mimo swoich niedoskonałości (które nie są dyskutowane w tej pracy) można wskazać na możliwość jej wykorzystania w procedurze wyznaczania podobieństwa dokumentów. Chcąc wyznaczyć podobieństwo pomiędzy dokumentami reprezentowanymi przez i -tą oraz j -tą kolumnę, można posłużyć się wzorem:

$$p_{XML}(D_i, D_j) = \frac{\sum_{k=1}^N x_{ki}^{bin} \cdot x_{kj}^{bin}}{N}, \quad (8)$$

w którym wartość N jest liczbą rozważanych wyrazów (a tym samym liczbą wierszy w macierzy częstości).

Zaproponowany miernik wyznacza liczbę wyrazów występujących w obu porównywanych dokumentach w stosunku do liczby wszystkich wyrazów wchodzących w skład przetwarzanego zbioru.

Jako przykład rozpatrzony zostanie problem wyznaczenia podobieństwa pomiędzy dwoma krótkimi dokumentami tekstowymi: Warto zauważyć, że przed realizacją właściwych obliczeń sprowadzono do formy podstawowej wszystkie wyrazy występujące w tekście.

Tekst 1: Mowa jest srebrem, lecz milczenie złotem.

Tekst 2: Milczenie bywa wymowniejsze od mowy.

Podobieństwo przyjmuje wartość:

$$p_{XML}(\text{Tekst 1}, \text{Tekst 2}) = 3/8.$$

5. Podobieństwo elementów złożonych

W odróżnieniu od elementów prostych, elementy złożone mają wartości reprezentowane za pomocą struktur drzewiastych. Wyznaczając

$$p_{XML}(\langle n_i, E_i \rangle, \langle n_j, E_j \rangle), \quad (9)$$

należy pamiętać, że podobieństwo E_i oraz E_j uzależnione jest od różnic w ich strukturze oraz od zróżnicowania przechowywanych w nich wartości.

Również w przypadku wyznaczania podobieństwa elementów złożonych rozpatrzeć należy kilka przypadków.

a) podobieństwo elementów złożonych o różnych nazwach

Jeżeli nazwy elementów są różne, to ich podobieństwo jest zerowe:

$$p_{XML}(\langle n_i, E_i \rangle, \langle n_j, E_j \rangle) = 0, \text{ gdy } n_i \neq n_j. \quad (10)$$

W celu przedstawienia sposobu korzystania z formuły wyznaczone zostanie podobieństwo elementów o nazwach TOWAR oraz PRODUKT.

$X1 = \langle \text{TOWAR} \rangle$

$\langle \text{NAZWA} \rangle \text{Chleb wiejski} \langle / \text{NAZWA} \rangle$

$\langle \text{PRODUCENT} \rangle \text{Piekarnia Tradycyjna} \langle / \text{PRODUCENT} \rangle$

$\langle \text{CENA} \rangle 1.70 \langle / \text{CENA} \rangle$

$\langle / \text{TOWAR} \rangle$

$X2 = \langle \text{PRODUKT} \rangle$

$\langle \text{NAZWA} \rangle \text{Zapałki} \langle / \text{NAZWA} \rangle$

$\langle \text{CENA} \rangle 0.35 \langle / \text{CENA} \rangle$

$\langle / \text{PRODUKT} \rangle,$

czyli:

$$p_{XML}(X1, X2) = 0.$$

b) podobieństwo elementów złożonych w przypadku wystąpienia wartości pustych

Jeżeli wartości w obu elementach są zbiorami pustymi, to miernik ich podobieństwa wynosi jeden.

$$p_{XML}(\langle n_i, \emptyset \rangle, \langle n_i, \emptyset \rangle) = 1. \quad (11)$$

Jeżeli zaś jeden element ma wartość w postaci zbioru pustego, a drugi zbioru niepustego, to ich podobieństwo jest zerowe:

$$p_{XML}(\langle n_i, E_i \rangle, \langle n_i, \emptyset \rangle) = p_{XML}(\langle n_i, \emptyset \rangle, \langle n_i, E_j \rangle) = 0. \quad (12)$$

c) podobieństwo elementów, z których jeden posiada wartość typu prostego, a drugi typu złożonego

W przypadku porównywania elementów, z których jeden posiada wartość typu prostego, a drugi typu złożonego, podobieństwo jest zerowe

$$p_{XML}(\langle n_i, w_i \rangle, \langle n_i, E_j \rangle) = 0. \quad (13)$$

Jako przykład opisywanej w bieżącym punkcie sytuacji można rozważyć:

```
X1 =      <TOWAR>
          <NAZWA>Chleb wiejski</NAZWA>
          <PRODUCENT>Piekarnia Tradycyjna</PRODUCENT>
          <CENA>1.70</CENA>
        </TOWAR>
X2 =      <TOWAR>Masło</TOWAR>,
```

po dokonaniu obliczeń:

$$p_{XML}(X1, X2) = 0.$$

d) podobieństwo obiektów złożonych

Z całą pewnością jest to najciekawszy przypadek w procedurze wyznaczania podobieństwa. Przyjmując, że wartości porównywanych elementów są zbiorami wartości prostych, wyznaczyć można macierz podobieństwa elementów złożonych. Macierz ta zawiera wartości mierników podobieństwa pomiędzy elementami składowymi zbiorów stanowiących wartości porównywanych elementów i przyjmuje postać:

$$S = \begin{bmatrix} p_{XML}(el_{i1}, el_{j1}) & p_{XML}(el_{i1}, el_{j2}) & p_{XML}(el_{i1}, el_{j|E_j|}) \\ p_{XML}(el_{i2}, el_{j1}) & p_{XML}(el_{i2}, el_{j2}) & p_{XML}(el_{i2}, el_{j|E_j|}) \\ \dots & \dots & \dots \\ p_{XML}(el_{i|E_i|}, el_{j1}) & p_{XML}(el_{i|E_i|}, el_{j2}) & p_{XML}(el_{i|E_i|}, el_{j|E_j|}) \end{bmatrix}. \quad (14)$$

Warto zauważyć, że liczba wierszy w macierzy S nie musi być równa liczbie kolumn. Po wyznaczeniu macierzy podobieństwa należy połączyć ze sobą elementy porównywanych zbiorów w taki sposób, aby maksymalizować ich łączne podobieństwo (należy dążyć do maksymalizacji sumy elementów znajdujących się na przecięciu łączonych ze sobą wierszy i kolumn). W realizacji praktycznej problem ten rozwiązywany jest zwykle za pomocą algorytmu węgierskiego.

Podobieństwo łączne elementów złożonych o wartościach E_i oraz E_j wyraża wzór

$$p_{XML}(\langle n_i, E_i \rangle, \langle n_j, E_j \rangle) = \frac{s}{|E_i \cup E_j|}, \quad (15)$$

gdzie s jest sumą elementów macierzy S znajdujących się na przecięciu łączonych wierszy i kolumn, $|E_i \cup E_j|$ zaś jest liczbą elementów w zbiorze będącym sumą zbiorów E_i oraz E_j .

Przykładowe obliczenia dotyczyć będą wyznaczenia podobieństwa elementów:

```
X1 = <TOWAR>
      <NAZWA>Chleb wiejski</NAZWA>
      <PRODUCENT>Piekarnia Tradycyjna</PRODUCENT>
      <CENA>1.70</CENA>           //cena max 5
</TOWAR>
```

oraz

```
X2 = <TOWAR>
      <NAZWA>Zapałki</NAZWA>
      <CENA>0.35</CENA>           //cena max 5.
</TOWAR>
```

Wyznaczona macierz podobieństwa przyjmuje postać:

$$S = \begin{bmatrix} & NAZWA & CENA \\ NAZWA & 0 & 0 \\ PRODUCENT & 0 & 0 \\ CENA & 0 & 0,881 \end{bmatrix},$$

a obliczony miernik podobieństwa wynosi:

$$p_{XML}(X1, X2) = 0,881/3 = 0,294.$$

Rozważając zagadnienie budowy macierzy podobieństwa S , należy zwrócić uwagę na to, że elementy porównywanych zbiorów E_i oraz E_j nie muszą być typu prostego. Jeśli są typu złożonego, to w celu obliczenia miernika podobieństwa pomiędzy nimi a elementami (prostymi lub złożonymi) drugiego zbioru należy w sposób rekurencyjny zastosować ten sam algorytm.

6. Wnioski końcowe

Przedstawione w pracy rozważania dotyczą wyznaczania podobieństwa pomiędzy obiektami reprezentowanymi za pomocą struktur drzewiastych i opisanymi w formacie XML. Metody wyznaczania tego typu podobieństwa, a szczególnie propozycja zaprezentowana w niniejszej pracy, mogą znaleźć zastosowanie przy rozwiązywaniu wszystkich tych problemów badawczych, w których niezbędne jest wyznaczanie podobieństwa pomiędzy obiektami (do tej grupy problemów należy zaliczyć problemy z zakresu klasyfikacji).

Jest rzeczą oczywistą, że przedstawiona propozycja wymaga dalszych badań oraz udoskonaleń. W przypadku potwierdzenia jej przydatności należy rozważyć możliwość jej rozbudowy poprzez umożliwienie przypisywania wag do poszczególnych elementów oraz uwzględniania wartości argumentów przypisanych do poszczególnych znaczników.

Literatura

- Bhavsar V., Boley H., Yang L., *A Weighted-Tree Similarity Algorithm, for Multi-Agent Systems in E-Business Environments*, „Computational Intelligence Journal” 2004 20(4).
- Cormen T.H., Leiserson C.E., Rivest R.L., Stein C., *Wprowadzenie do algorytmów*, Wydawnictwa Naukowo-Techniczne, Warszawa 2004.
- Long J., Schwartz D., Stoecklin S., *An XML Distance Measure*, 2005, <http://ww2.cs.fsu.edu/~jidl原因/publications/dmin-camera.pdf>.

THE SIMILARITY OF OBJECTS REPRESENTED BY TREES

Summary

The Extensible Markup Language (XML) belongs to the group of the most important and the most popular tools of contemporary computer science. The XML may be used for the description of object structure and for data gathering. The scope of the XML applications is very wide and it spreads from the problems of gathering and processing multimedia data, through the database solutions, to the designing and implementation of web services.

Trees are the most important data structure in XML language. The XML tree consists of nodes and edges. The nodes may include quantitative, qualitative or textual data. The significance of edges may be expressed by weights. The flexibility of the trees causes that they can be considered as a suitable tool for description of complex objects and hierarchical relationships between elements. On the other hand the processing of tree data is very difficult.

The calculation of the similarity measure between XML trees is the main issue of the paper. The survey of propositions on that subject may be found in the first part of the article. Next a full procedure of the calculation of the tree similarity measure is presented. The final remarks and the directions of further investigations are presented in the last part of the paper.