

Paweł Lula

Akademia Ekonomiczna w Krakowie

CZYNNIKI DETERMINUJĄCE WYZNACZANIE ODLEGŁOŚCI POMIĘDZY DRZEWAMI ZDEFINIOWANYMI W JĘZYKU XML

1. Wstęp

Tematyka niniejszego opracowania skupiona jest wokół problemu wyznaczania odległości pomiędzy obiektami opisywanymi za pomocą struktur drzewiastych. Zakres omawianych problemów ograniczony został do tych aspektów, które znajdują swoje zastosowanie przy wyznaczaniu odległości pomiędzy obiektami zdefiniowanymi w postaci zapisów w języku XML.

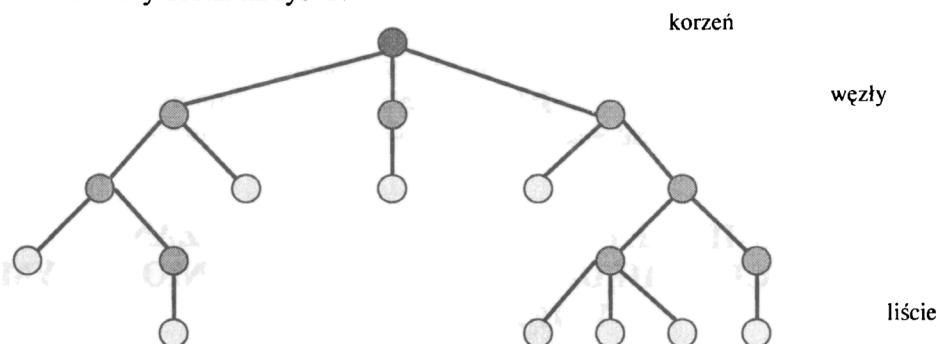
Podstawowymi celami pracy są identyfikacja czynników determinujących odległość pomiędzy drzewami zdefiniowanymi za pomocą języka XML oraz wskazanie ich wpływu na wyznaczaną odległość.

W kolejnych punktach pracy zamieszczono charakterystykę struktur drzewiastych oraz zaprezentowano ich rolę w języku XML. Następnie dokonano krótkiego przeglądu znanych z literatury propozycji metod pomiaru odległości pomiędzy drzewami. Dalsze punkty pracy zawierają próbę identyfikacji elementów, których uwzględnienie jest konieczne przy formułowaniu miary odległości dla drzew. Końcowe fragmenty opracowania zawierają wnioski podsumowujące oraz uwagi implementacyjne dotyczące praktycznych aspektów realizacji obliczeń.

2. Struktury drzewiaste i ich rola w technologii XML

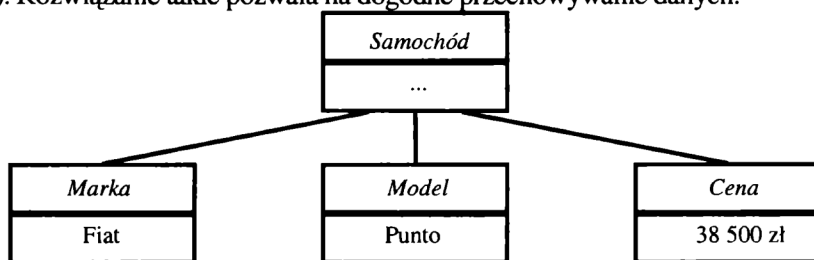
Drzewa należą do grupy najważniejszych i najpopularniejszych złożonych struktur danych. Posługując się pojęciami zaczerpniętymi z teorii grafów, drzewo definiowane jest jako spójny, acykliczny graf nieskierowany [Cormen i in. 2004].

W sposób opisowy, na użytek zastosowań praktycznych, drzewo definiowane jako struktura danych złożona z połączonych ze sobą w sposób hierarchiczny elementów składowych zwanych węzłami. Na szczycie hierarchii znajduje się element początkowy określany mianem korzenia. Węzły posiadające elementy potomne nazywane są węzłami wewnętrznymi, natomiast węzły nieposiadające potomków to liście (porównaj: <http://www.nist.gov/dads/HTML/tree.html>). Schemat drzewa przedstawiony został na rys. 1.



Rys. 1. Przykładowe drzewo
Źródło: opracowanie własne.

W ostatnich latach znaczenie i popularność struktur drzewiastych wzrosły wraz z rozwojem języka XML (*extensible markup language* – rozszerzalny język znaczników, 1997), w którym drzewa stanowią podstawową strukturę danych. Są to drzewa etykietowane, w których z poszczególnymi węzłami mogą być skojarzone etykiety oraz wartości (rys. 2). Rozwiązanie takie pozwala na dogodne przechowywanie danych.



Rys. 2. Przykładowe drzewo z etykietowanymi węzłami (etykiety węzłów zapisano kursywą, ich wartości zaś – czcionką prostą)
Źródło: opracowanie własne.

Technologia XML jest doskonałym narzędziem do tworzenia specjalizowanych języków służących do definiowania struktur danych specyficznych dla wybranej dziedziny zastosowań. Umożliwia również przechowywanie danych i sprawdzania ich poprawności względem przyjętej wcześniej definicji oraz może służyć do definiowania sposobu ich przetwarzania. Ciekawsze zastosowania technologii XML obejmują m.in.:

- definiowanie struktury i przechowywanie danych prezentowanych w serwisach internetowych;
- reprezentację dokumentów wymienianych drogą elektroniczną;
- zapewnienie możliwości integracji systemów komputerowych przez zastosowanie formatu XML jako wspólnego narzędzia komunikacji dla komunikujących się ze sobą systemów;
- opis grafiki wektorowej (SVG – *scalable vector graphics*);
- narzędzie definiowania formuł matematycznych (MathML – *mathematical markup language*);
- definiowanie postaci modeli predykcyjnych (PMML – *predictive model markup language*), pozwalające na bezproblemową wymianę definicji modeli między różnymi pakietami obliczeniowymi;
- narzędzia tworzenia serwisów dostępnych za pośrednictwem systemów mobilnych (*wireless markup language* – WML);
- języki przeznaczone do opisu danych rozpatrywanych w poszczególnych dziedzinach wiedzy; za przykład mogą posłużyć:
 - CML – *chemical markup language*,
 - BSML – *bioinformatic sequence markup language*,
 - AML – *astronomical markup language*,
 - ThML – *theological markup language*.

Ciągle rozszerzanie zakresu zastosowań rozszerzalnego języka znaczników powoduje, że zasoby informacyjne przechowywane w postaci dokumentów XML bardzo szybko rosną. Oczywistą konsekwencją tego faktu jest zwiększanie się potrzeby przystosowania metod analizy danych do operowania na tego typu strukturach. Realizacja takich obliczeń wymaga określenia sposobu operowania na strukturach drzewiastych w sposób zapewniający uwzględnienie zarówno informacji wpływających ze struktury drzewa, jak również wartości przechowywanych w jego węzłach. Jedną z najważniejszych operacji przeprowadzanych w trakcie analizy danych jest wyznaczenie odległości pomiędzy porównywanymi obiektami. Operacja ta umożliwia m.in. analizę taksonomiczną rozpatrywanego zbioru elementów.

3. Propozycje metod pomiaru odległości pomiędzy drzewami XML

Podstawową metodą określania odległości uwzględniającej zróżnicowanie strukturalne jest odległość edycyjna. Określa ona minimalną liczbę operacji elementarnych koniecznych do przekształcenia jednej struktury w drugą. Idea pomiaru odległości edycyjnej pochodzi od Wagnera i Fishera [Wagner, Fisher 1974], którzy wykorzystali ją do określania odległości pomiędzy łańcuchami znaków. Wyróżnili oni trzy operacje elementarne: zastąpienie znaku innym znakiem, usunięcie znaku oraz wstawienie znaku. Odległość pomiędzy łańcuchami zdefiniowana została jako minimalna liczba wymienionych operacji potrzebnych do prze-

kształcenia jednego łańcucha w drugi. W roku 1977 Selkow zaproponował wykorzystanie analogicznej metody do pomiaru odległości pomiędzy drzewami [Selkow 1977]. Selkow zdefiniował następujące operacje elementarne: zmianę etykiety, usunięcie oraz wstawienie węzła, przy czym dwie ostatnie operacje dotyczyć mogły wyłącznie liści. W kolejnych latach wielu innych autorów proponowało również własne modyfikacje odległości edycyjnej wyrażającej zróżnicowanie między drzewami. Do najbardziej znanych należy zaliczyć metody zaproponowane w pracach [Tai 1979] oraz [Zhang, Shasha 1989]. W kontekście struktur drzewiastych wykorzystywanych w języku XML żadna z powyższych metod nie rozwiązuje problemu pomiaru odległości w sposób całościowy, gdyż są one przystosowane do operowania na węzłach posiadających wyłącznie przypisane etykiety i nie znajdują zastosowania w przypadku charakterystycznych dla technologii XML węzłów posiadających zarówno etykiety, jak i wartości.

Propozycje rozwiązania problemu pomiaru odległości pomiędzy drzewami charakterystycznymi dla języka XML można znaleźć w pracach [Bhavsar, Boley, Yang 2004] oraz [Long, Schwartz, Stoecklin 2005]. Analiza proponowanych algorytmów pozwala stwierdzić, że każdy z nich może służyć do wyrażenia odległości pomiędzy drzewami języka XML. Trzeba jednak zauważyć, że konstrukcja obu mierników jest odmienna i różne są uzyskiwane wyniki. Porównując oba podejścia, niezwykle trudno jest wskazać metodę lepszą. Z całą pewnością sposób oceny zależy od rozpatrywanego problemu i założeń przyjętych przez badacza. Wydaje się, że również próby oceny jakości poszczególnych miar odległości dokonane na podstawie eksperymentów symulacyjnych nie doprowadzą do jednoznacznego rozwiązania problemu ze względu na brak jednego, obiektywnego kryterium określania jakości badanych miar. Problem określenia metody pomiaru odległości pomiędzy drzewami należy raczej rozpatrywać indywidualnie dla każdego rozważanego problemu badawczego. Aby zrealizować ten proces w sposób prawidłowy, należy zidentyfikować wszystkie czynniki wpływające na odległość pomiędzy drzewami, zaproponować metody ich pomiaru oraz agregacji uzyskanych w ten sposób mierników cząstkowych w zagregowaną miarę odległości.

4. Czynniki determinujące odległość pomiędzy drzewami

Zasadniczym celem tego punktu jest przedstawienie propozycji listy czynników, które należy uwzględnić przy wyrażaniu odległości pomiędzy drzewami.

4.1. Sposób pomiaru odległości pomiędzy wartościami przechowywanymi w liściach drzew

Język XML dopuszcza możliwość przechowywania w liściach drzew wartości liczbowych, łańcuchów znaków i tekstów. W szczególnym przypadku liść drzewa

może przechowywać odwołanie do pliku dyskowego, co pozwala na umieszczenie w strukturze drzewa danych graficznych, filmów oraz dźwięków. Typ wartości dwóch porównywanych elementów determinuje sposób wyrażenia odległości pomiędzy nimi. Dla danych wyrażonych na klasycznych skalach pomiarowych sformułowanie sposobu pomiaru odległości nie stwarza zwykle trudności. Źródłem problemów mogą być dane tekstowe oraz dane multimedialne, których prawidłowe uwzględnienie w obliczeniach wymaga jeszcze wielu badań ukierunkowanych przede wszystkim na wypracowanie metod pozwalających na uwzględnienie semantycznych aspektów porównywanych wartości. Decyzje dotyczące sposobu operowania na liściach muszą również obejmować kwestię unormowania tworzonych mierników.

Warto zauważyć, że przywołane propozycje miar odległości dla obiektów języka XML w zakresie operowania na poszczególnych wartościach są dalekie od doskonałości, gdyż miara zaproponowana w pracy [Long, Schwartz, Stoecklin 2005] przystosowana jest jedynie do działania na wartościach nominalnych i numerycznych. Autorzy nie uwzględniają możliwości wystąpienia danych innego typu. Wartość miary odległości pomiędzy elementami unormowana jest do przedziału [0; 1]. Propozycja zawarta w pracy [Bhavsar, Boley, Yang 2004] jest jeszcze bardziej uboga, gdyż uwzględnia możliwość umieszczenia w liściach drzew jedynie wartości nominalnych. Propozycje pomiaru odległości pomiędzy węzłami przechowującymi teksty znaleźć można w pracy [Corley, Csomai, Mihalcea 2005].

4.2. Sposób traktowania atrybutów elementów

Specyfikacja języka XML pozwala na umieszczanie atrybutów w poszczególnych węzłach. Za przykład posłużyć mogą następujące zapisy:

```
<PALIWO CENA="4,05">Olej napędowy</PALIWO>
```

```
<PALIWO DOSTAWCA="Orlen">Olej napędowy</PALIWO>
```

Atrybuty służą zwykle do wzbogacenia informacji reprezentowanej przez węzeł o dodatkowe elementy. Sposób interpretacji atrybutów elementów jest specyficzną cechą rozważanego problemu i w sposób zindywidualizowany należy podejmować decyzje o potrzebie i sposobie ich uwzględnienia przy obliczaniu odległości pomiędzy elementami. Autorzy pracy [Long, Schwartz, Stoecklin 2005] uwzględniają możliwość wystąpienia atrybutów. Niestety proponowane podejście nie jest elastyczne, gdyż przewiduje uwzględnienie wszystkich atrybutów, niezależnie od sposobu ich interpretacji. Trudno również zaakceptować pogląd autorów głoszący, że różnice w wartościach atrybutów mają identyczne znaczenie jak różnice w wartościach elementów. Natomiast propozycja w pracy [Bhavsar, Boley, Yang 2004] pomija całkowicie możliwość wystąpienia atrybutów.

4.3. Formuła agregacji odległości pomiędzy elementami składowymi

Przy wyznaczaniu odległości pomiędzy dwoma węzłami wewnętrznymi należy przeprowadzać agregację odległości wyznaczonych dla elementów potomnych obu węzłów. Decyzja o wyborze właściwej formuły powinna zostać podjęta na podstawie analizy merytorycznej rozpatrywanego problemu. W przypadku miernika zaproponowanego w [Long, Schwartz, Stoecklin 2005] agregacja dokonywana jest poprzez obliczenie średniej arytmetycznej. Natomiast autorzy pracy [Bhavsar, Boley, Yang 2004] zdecydowali się na zastosowanie średniej ważonej. Zaletą drugiego rozwiązania jest możliwość zróżnicowania ważności poszczególnych elementów potomnych, problemem natomiast może być ustalanie właściwych wartości wag.

4.4. Problem ważenia elementów potomnych

Poruszona kwestia ważenia znaczenia elementów potomnych jest bardzo trudna do rozstrzygnięcia. Przypisanie identycznych wag poszczególnym potomkom nie zawsze jest zgodne z ich rzeczywistym znaczeniem. Określenie wag definiujących ich istotność wymaga w wielu przypadkach wiedzy eksperckiej i może mieć charakter subiektywny. W rozpatrywanych w pracy algorytmach za stosowaniem identycznych wag opowiedzieli się autorzy pracy [Long, Schwartz, Stoecklin 2005]. Natomiast twórcy propozycji wyrażonej w pracy [Bhavsar, Boley, Yang 2004] opowiadają się za możliwością ważenia znaczenia poszczególnych elementów potomnych. Przyjęli oni, że wagi określane są przez osobę prowadzącą obliczenia. W celu uproszczenia obliczeń postulują oni przypisanie potomkom węzła takich wartości współczynników wagowych, których suma równa się jedności. Oprócz tych typowych cech systemu wag omawiana grupa autorów dopuściła również możliwość zastosowania różnych wartości współczynników wagowych w obu porównywanych drzewach.

4.5. Sposób uwzględnienia informacji o zależnościach hierarchicznych

Ważność elementów drzewa jest uzależniona od ich położenia. Znaczenie elementu maleje wraz ze zwiększaniem się jego głębokości. Ta intuicyjnie odczuwana prawidłowość jest uwzględniona również w badanych algorytmach. Zarówno w pracy [Long, Schwartz, Stoecklin 2005], jak i w artykule [Bhavsar, Boley, Yang 2004] znaczenie różnicy pomiędzy elementami jest odwrotnie proporcjonalne do głębokości ich wystąpienia w porównywanych drzewach. Jednakże w rozwiązaniu z pracy [Bhavsar, Boley, Yang 2004] zastosowano dodatkowo prosty, ale bardzo przydatny mechanizm definiowania relatywnej ważności zgodności typu porównywanych węzłów wewnętrznych w stosunku do ważności zgodności elementów potomnych. Efekt ten uzyskano przez wprowadzenie parametru algorytmu w postaci współczynnika z przedziału $[0; 1]$ określającego wspomnianą powyżej relację.

4.6. Niezgodności w strukturach porównywanych drzew

Omawiany w tym miejscu problem pojawia się wówczas, gdy elementy potomne porównywanych ze sobą węzłów wewnętrznych są różnego typu. Ze względu na możliwość występowania wśród potomków elementów będących węzłami rodzicielskimi dla kolejnych składników drzewa wspomniana niezgodność typów potomków może prowadzić do całkowitego zróżnicowania struktur rozpatrywanych obiektów. Zespół autorski pracy [Long, Schwartz, Stoecklin 2005] przyjął w tym zakresie stosunkowo proste, ale niepozbawione uzasadnienia rozwiązanie: odległość pomiędzy elementami różnych typów wynosi jeden (ze względu na unormowanie miary wartość ta oznacza odległość maksymalną). Autorzy pracy [Bhavsar, Boley, Yang 2004] proponują zastosowanie procedury określającej stopień podobieństwa elementu (liścia lub poddrzewa) do pustego (brakującego) elementu w drugim obiekcie. Podobieństwo to zmniejsza się wraz ze zwiększaniem się szerokości i głębokości drzewa niemającego swojego odpowiednika w drugim obiekcie.

5. Wnioski końcowe

Celem przedstawionych rozważań była próba przedstawienia listy czynników koniecznych przy wyznaczaniu odległości pomiędzy drzewami wraz z pokazaniem sposobu podejścia do każdego z tych problemów przez autorów dwóch najbardziej znanych algorytmów. Proponowane metody postępowania należy uznać za bardzo interesujące, ale z całą pewnością nie mają one charakteru uniwersalnego, przystosowanego do każdego z możliwych problemów. Najprawdopodobniej w praktyce sformułowanie jedyne, zawsze poprawnego podejścia nie jest możliwe. Należy raczej przyjąć, że sposób postępowania zmierzający do wyznaczenia odległości pomiędzy drzewami powinien być ustalany w sposób zindywidualizowany dla każdego problemu. Przyjęcie takiej strategii postępowania pociąga za sobą konieczność dostosowywania do każdego zagadnienia oprogramowania służącego do obliczeń. Chcąc ułatwić proces modyfikacji narzędzi programowych, należy preferować takie rozwiązania, które z jednej strony charakteryzują się wysokim stopniem elastyczności w zakresie operowania na drzewach, z drugiej zaś strony są stosunkowo łatwe w użyciu. Doświadczenia autora niniejszego tekstu związane z tworzeniem narzędzi programowych służących do operowania na strukturach w formacie XML wskazują na szczególną przydatność obiektowego modelu dokumentu (modelu DOM – *document object model*) dostarczającego niezależnego od języka programowania mechanizmu operowania strukturach drzewiastych języka XML.

Literatura

- Bhavsar V.C., Boley H., Yang L. (2004), *A Weighted-Tree Similarity Algorithm for Multi-Agent Systems in E-business Environments*, <http://www.cs.unb.ca/~boley/treesimilarity/BhavsarBoleyYang-CI.pdf>, 15.10.2006.
- Corley C., Csomai A., Mihalcea R. (2005), *Text Semantic Similarity with Applications*, Department of Computer Science, University of North Texas, <http://cerl.unt.edu/~corley/pub/paper65.ranlp.2005.pdf>, 15.10.2006.
- Cormen T.H., Leiserson C.E., Stein C. (2004), *Wprowadzenie do algorytmów*, Wydawnictwo Naukowo-Techniczne, Warszawa.
- Long J., Schwartz D.G., Stoecklin S. (2005), *An XML Distance Measure*, <http://www.cs.fsu.edu/~stoeckli/PublicationsVitae/XMLDistanceMeasure.pdf>, 15.10.2006.
- Selkow S.M. (1977), *The Tree-to-Tree Editing Problem*, „Information Processing Letters” vol. 6, nr 6, s. 184-187.
- Tai K.-C. (1979), *The Tree-to-Tree Correction Problem*, „J. ACM” vol. 26, nr 3, s. 422-433.
- Wagner R., Fisher M. (1974), *The String-to-String Correction Problem*, „J. ACM” vol. 21, nr 1, s. 168-173.
- Zhang K., Shasha D. (1989), *Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems*, „SIAM Journal on Computing” vol. 18, nr 6, s. 1245-1262.

THE DETERMINANTS OF THE DISTANCE BETWEEN XML TREES

Summary

The issue of distance calculation between XML trees is the main subject of the paper. Tree structures and their applications in XML technology were shown in the first part of the study. Some remarks on edit distance between trees and on two new algorithms of XML distance calculation are presented next. Finally the list of factors which determine the distance between trees is proposed and discussed. The analysis shows that the universal method of distance calculation does not exist and it is necessary to formulate the method of tree distance evaluation for each problem independently.