STUDIA INFORMATICA 2009
Volume 30 Number 2A (83)

Daniel KOSTRZEWA, Henryk JOSIŃSKI
Politechnika Śląska, Instytut Informatyki

ZASTOSOWANIE ALGORYTMU IWO DO PLANOWANIA PROCESU SCALANIA DANYCH ROZPROSZONYCH

Streszczenie. Prezentowane zagadnienie stanowi kontynuację badań poświęconych zastosowaniu algorytmu ewolucyjnego do realizacji zadania istotnego dla dziedziny rozproszonych baz danych – określenia planu przebiegu procesu scalania danych rozproszonych. Wskazano zarówno na cechy wspólne, jak i różnice między algorytmami IWO i ewolucyjnym, zamieszczono rezultaty eksperymentów porównawczych.

Słowa kluczowe: algorytm IWO, optymalizacja zapytań, rozproszona baza danych

APPLICATION OF THE INVASIVE WEED OPTIMIZATION ALGORITHM FOR PREDETERMINATION OF THE PROGRESS OF DISTRIBUTED DATA MERGING PROCESS

Summary. The considered issue is a continuation of research concerning the application of the evolutionary algorithm for realization of the important task from the domain of distributed databases – predetermination of the progress of distributed data merging process. Many common features of the IWO method and the evolutionary algorithm as well as the differences between them were mentioned in the paper along with results of comparative experiments.

Keywords: the IWO algorithm, query optimization, distributed database

1. Wstęp

Rozpatrywane zagadnienie stanowi kontynuację badań poświęconych zastosowaniu algorytmu ewolucyjnego do realizacji tego samego zadania, istotnego dla dziedziny rozproszonych baz danych – określenia planu przebiegu procesu scalania danych rozproszonych.

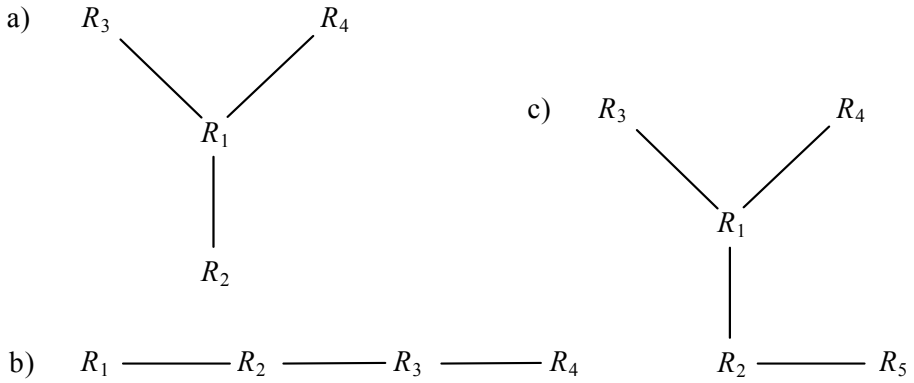
Rezultaty tych badań opisano w pracy [3]. Wybór algorytmu IWO jako kontrpropozycji względem algorytmu ewolucyjnego nie był przypadkowy – obie metody posiadają wiele cech wspólnych, do których zaliczyć należy: interpretację rozwiązania jako osobnika pewnej populacji, działania modyfikujące postać osobnika, model obliczeniowy dla funkcji przystosowania osobnika oraz obserwację wielu kolejno po sobie następujących populacji. Z drugiej strony, różnice między nimi, wyrażające się np. brakiem w algorytmie IWO operatora stanowiącego odpowiednik operatora krzyżowania w algorytmie ewolucyjnym oraz w odmiennych strategiach selekcji osobników do nowej populacji, sprawiają, że porównanie rezultatów działania obu algorytmów zastosowanych do rozwiązania tego samego problemu może przynieść interesujące spostrzeżenia.

2. Sformułowanie problemu

Proces scalania danych rozproszonych w zbiór rekordów stanowiący *wynik końcowy* zapytania adresowanego do rozproszonej bazy danych przebiega wieloetapowo i polega na wykonaniu wielu *operacji scalających*. Celem pojedynczej operacji jest scalenie dwóch zbiorów rekordów, z których każdy może być zbiorem pobranym z jednej z baz danych tworzących bazę rozproszoną (*zbiorem pierwotnym*) lub może stanowić wynik operacji scalającej przeprowadzonej w jednym z wcześniejszych etapów (*wynik pośredni*). Operację scalającą można wykonać w dowolnym węźle, w którym działa program ją realizujący. Spośród kilku typów operacji scalających (złączenie, złączenie zewnętrzne, suma zbiorów rekordów) najczęściej w realizacji zapytań występuje złączenie [10]. Stąd celem badań stało się skonstruowanie algorytmu, który posłuży do wyznaczenia takiego porządku i miejsc realizacji złączeń, aby wynik końcowy dotarł do węzła *docelowego* (miejsca, w którym użytkownik oczekuje na dane) w jak najkrótszym czasie. Przestrzenią poszukiwań dla rozpatrywanego zagadnienia jest zbiór planów realizacji zadanego zapytania reprezentowanego przez *graf złączeń* (ang. *join graph*). Graf złączeń stanowi strukturę złożoną z wierzchołków połączonych nieskierowanymi krawędziami. Wierzchołki grafu są odpowiednikami zbiorów pierwotnych. Każda z krawędzi grafu wyraża fakt istnienia między dwoma zbiorami *wyrażenia łączącego*, czyli kryterium dopasowania rekordów tych zbiorów do siebie. Do charakterystycznych kształtów grafu złączeń zalicza się:

- graf *gwiazdy*, w którym jeden ze zbiorów rekordów występuje w każdym wyrażeniu łączącym (jest centralnym wierzchołkiem grafu przedstawionego na rys. 1a),
- graf *łańcuchowy*, gdzie dwa zbiory rekordów (stanowiące skrajne wierzchołki grafu przedstawionego na rys. 1b) występują tylko jednokrotnie w wyrażeniach łączących, natomiast pozostałe zbiory – dwukrotnie.

Pozostałe kształty określa się mianem grafów *nieregularnych* (rys. 1c).



Rys. 1. Graf złączeń: a) graf gwiazdy, b) graf łańcuchowy, c) graf nieregularny
Fig. 1. Shape of a join graph: a) star, b) chain, c) irregular

Strategie wyznaczające kolejność realizacji złączeń posługują się operacją ekspansji lub transformacji. Wielokrotnie zastosowana *ekspansja* pozwala konsekwentnie zbudować pełne rozwiązanie problemu przez znajdowanie kolejnych jego elementów. Natomiast *transformacja* przekształca pełne rozwiązanie w inne [4]. Na ekspansji opierają działanie metody deterministyczne, które dla danego grafu złączeń i określonej przestrzeni poszukiwań zawsze prowadzą do otrzymania takiego samego rozwiązania. Należą do nich: algorytm oparty na idei programowania dynamicznego oraz zachłanny. Do grupy metod posługujących się transformacją zaliczane są natomiast metaheurystyki kombinatoryczne (algorytm symulowanego wyżarzania, algorytm iteracyjnego poprawiania, algorytm z listą tabu) oraz algorytmy ewolucyjne. Losowy sposób przekształcania rozwiązania w nowe decyduje o niedeterministycznym charakterze tych metod.

Wymienione algorytmy mogą służyć do wyznaczenia kolejności realizacji złączeń zarówno w scentralizowanej, jak i rozproszonej bazie danych. Aspekt rozproszenia danych wymaga jednak dodatkowo wyznaczenia miejsc realizacji poszczególnych operacji, co zwiększa stopień złożoności problemu.

Przegląd literatury opisującej metody rozwiązania omawianego zagadnienia byłby niezwykle obszerny. Pozycje związane z zastosowaniem algorytmu ewolucyjnego wymieniono w pracy [3]. Natomiast algorytm IWO, według wiedzy Autorów, nie został wcześniej użyty do wyznaczenia kolejności realizacji złączeń.

3. Idea algorytmu IWO

Algorytm IWO (ang. *Invasive Weed Optimization*) jest metodą optymalizacji, w której technika penetracji przestrzeni poszukiwań, oparta na transformacji, została zainspirowana

obserwacją cech charakterystycznych dla chwastów – ich gwałtownego rozprzestrzeniania się oraz szybkiego przystosowywania się do warunków otoczenia [5, 6, 9].

Wyróżnić należy następujące zasadnicze etapy algorytmu IWO:

1. Rozsianie nasion pierwszej populacji roślin.
2. Reprodukcja – wzrost chwastów i rozsiewanie nasion wokół siebie.
3. Selekcja najlepiej przystosowanych roślin w celu stworzenia z nich kolejnej populacji.

Pierwszy etap algorytmu polega na losowym rozsianiu nasion chwastów w przestrzeni poszukiwań, tym samym odpowiadając losowemu wygenerowaniu określonej liczby rozwiązań rozpatrywanego problemu. Liczebność pierwszej populacji chwastów jest jednym z parametrów algorytmu.

Liczba nasion rozsiewanych przez pojedynczą roślinę w etapie drugim zależy od stopnia jej przystosowania do środowiska – im większa wartość funkcji przystosowania danego chwastu, tym większa liczba nasion z niego pochodzących. Zależność ta ma charakter liniowy. Odległość między miejscem upadku nasienia a rośliną macierzystą opisana jest przez rozkład normalny o średniej 0 i odchyleniu standardowym obciętym do wartości nieujemnych i zmniejszającym się wraz z każdą kolejną iteracją algorytmu, zgodnie z następującą formułą podaną w pracach [5, 6, 9]:

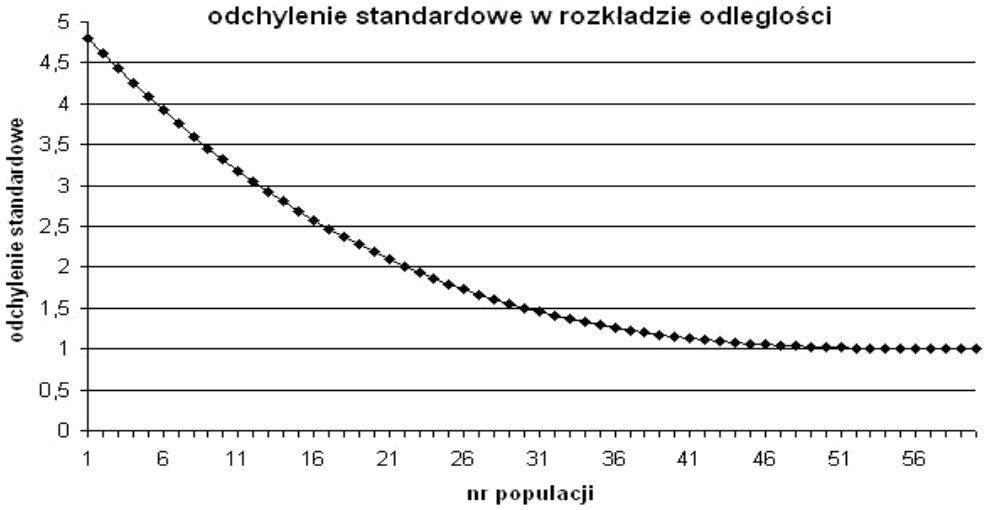
$$\sigma_{iter} = \left(\frac{iter_{\max} - iter}{iter_{\max}} \right)^m (\sigma_{init} - \sigma_{fin}) + \sigma_{fin} \tag{1}$$

Symbol σ_{iter} oznacza odchylenie standardowe obliczone dla iteracji o numerze *iter*. W każdej iteracji rozpatruje się kolejną populację chwastów, zatem łączna liczba iteracji ($iter_{max}$) równoważna jest łącznej liczbie populacji. Symbole σ_{init}, σ_{fin} reprezentują, odpowiednio, początkową i końcową wartość odchylenia standardowego, natomiast m jest współczynnikiem modulacji nieliniowej. Przebieg zmian wartości odchylenia standardowego zgodnie z formułą (1) dla wartości $iter_{max} = 60$, $m = 3$, $\sigma_{init} = 5$ oraz $\sigma_{fin} = 1$ przedstawia rys. 2.

Odległość między rośliną macierzystą a nowym chwastem wyraża stopień różnicy między nimi. Efekt owej różnicy uzyskuje się konstruując chwast potomny jako odpowiednik rośliny macierzystej poddanej wielokrotnej transformacji, przy czym liczba przeprowadzonych transformacji odpowiada odległości między obiema roślinami.

Ponieważ w drugim etapie algorytmu chwasty mogą rozsiewać znaczne liczby nasion, etap trzeci związany jest z koniecznością usunięcia części roślin. Eliminowane są chwasty najsłabiej przystosowane do środowiska w liczbie zapewniającej stałą liczebność kolejnych populacji.

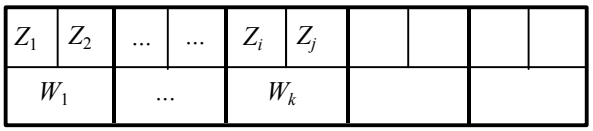
Etapy drugi i trzeci są wielokrotnie powtarzane. Liczba przeprowadzonych iteracji oraz inne wielkości występujące w formule (1) stanowią parametry algorytmu.



Rys. 2. Przebieg zmian wartości odchylenia standardowego dla kolejnych populacji
Fig. 2. Course of changes of the standard deviation for the consecutive populations

3.1. Reprezentacja rozwiązania

Pojedyncza roślina (osobnik populacji chwastów) stanowi odpowiednik pojedynczego rozwiązania problemu, którym jest określony porządek i zestaw miejsc realizacji złączeń. Jego symboliczną reprezentację przedstawia rys. 3.



Rys. 3. Reprezentacja pojedynczego rozwiązania
Fig. 3. Representation of a solution

Symbole Z_i, Z_j oznaczają numery zbiorów, które zostaną złączone w węźle o numerze W_k. Trójka symboli (Z_i, Z_j, W_k) będzie nazywana *genem*. Liczba genów w osobniku wynosi $n-1$, gdzie n jest liczbą zbiorów pierwotnych wymienionych w zapytaniu (zbiory pierwotne otrzymują numery od 1 do n). Układ genów odpowiada kolejności wykonywania poszczególnych złączeń. Wynik złączenia opisanego pierwszym genem będzie zbiorem rekordów o numerze $n+1$, a kolejne zbiory (wyniki pośrednie) otrzymają numery będące kolejnymi liczbami naturalnymi. Ponieważ wyniki pośrednie są używane w dalszych złączeniach, będą występować w kolejnych genach. Jednakże spełniony musi być warunek, że gen zawierający numer zbioru stanowiącego wynik pośredni może w osobniku wystąpić dopiero po genie, który reprezentuje operację złączenia generującą ten wynik pośredni.

3.2. Transformacja chwastu

Transformacja rośliny macierzystej, prowadząca do powstania chwastu potomnego dotyczy pojedynczego genu i polega na zmianie w składzie pary łączonych ze sobą zbiorów lub

na zmianie węzła, w którym ma odbyć się złączenie. Wartości prawdopodobieństwa transformacji numeru zbioru oraz numeru węzła stanowią parametry przeprowadzonych badań. Mechanizm wykorzystany w algorytmie IWO odpowiada operatorowi genetycznemu mutacji, zastosowanemu w algorytmie ewolucyjnym opisanym w pracy [3].

Transformacja numeru węzła polega na wylosowaniu wartości z puli dopuszczalnych numerów węzłów i zastąpieniu przez nią dotychczasowego numeru węzła w losowo wybranym genie.

Transformacja numeru zbiorów jest realizowana przez zamianę dwóch wylosowanych numerów zbiorów należących do różnych genów. W ten sposób nie dojdzie do niedopuszczalnego wielokrotnego wystąpienia tego samego numeru zbioru w pojedynczym osobniku. Ponadto, sposób przeprowadzenia transformacji zapewnia, że do genu trafia taki numer zbioru, dla którego gen, reprezentujący operację złączenia tworzącą ten zbiór, zajmuje w osobniku pozycję wcześniejszą od genu poddawanego transformacji.

3.3. Model obliczeniowy dla funkcji celu

Rozwiązanie zadania optymalizacji wymaga zdefiniowania funkcji celu, którą w rozważanym problemie jest szacunkowy koszt realizacji ciągu złączeń. Dla algorytmu IWO poddano modyfikacji model zastosowany w pracach [2, 3]. Wymaga on zdefiniowania formuł umożliwiających oszacowanie liczebności zbioru wynikowego operacji złączenia i rozmiaru rekordu w zbiorze wynikowym oraz kosztu operacji złączenia i operacji przesyłu danych.

Rozmiar rekordu w zbiorze wynikowym złączenia wyznaczany jest zgodnie z formułą:

$$dr(X \triangleright\triangleleft Y) = \eta(dr(X) + dr(Y)) \tag{2}$$

Symbole $dr(X)$, $dr(Y)$ oznaczają długości rekordów składowych, natomiast η jest *współczynnikiem redukcji*. Jeśli dla operacji złączenia dane jest wyrażenie łączące, to wartość współczynnika η pochodzi z przedziału (0, 1) i jest jednym z parametrów przeprowadzonych badań. W przeciwnym przypadku złączenie realizowane jest jako iloczyn kartezjański obu zbiorów, a współczynnik redukcji przyjmuje wartość 1.

Sposób wyznaczenia liczebności zbioru wynikowego operacji złączenia wymaga rozważenia następujących przypadków:

- jeśli nie podano wyrażenia łączącego, to liczebność zbioru wynikowego jako iloczynu kartezjańskiego obu zbiorów jest równa iloczynowi liczebności łączonych zbiorów $lr(X)$, $lr(Y)$:

$$lr(X \triangleright\triangleleft Y) = lr(X) \cdot lr(Y) \tag{3}$$

- jeśli między wartościami atrybutów obu zbiorów występujących w wyrażeniu łączącym występuje zależność polegająca na tym, że pojedynczej wartości atrybutu zbioru o nie-

większej liczebności odpowiada co najwyżej jedna wartość atrybutu drugiego zbioru, to poprawnym oszacowaniem liczebności zbioru wynikowego jest mniejsza spośród liczebności obu zbiorów (przykładem ilustrującym tę zależność może być wyrażenie łączące, w którym występują klucze główne obu zbiorów):

$$lr(X \triangleright\triangleleft Y) = \min(lr(X), lr(Y)) \tag{4}$$

- jeśli między wartościami atrybutów obu zbiorów występujących w wyrażeniu łączącym występuje zależność polegająca na tym, że pojedynczej wartości atrybutu zbioru X odpowiada wiele wartości atrybutu zbioru Y i równocześnie pojedynczej wartości atrybutu zbioru Y odpowiada co najwyżej jedna wartość atrybutu zbioru X, to poprawnym oszacowaniem liczebności zbioru wynikowego jest liczebność zbioru Y (przykładem ilustrującym tę zależność może być wyrażenie łączące, w którym występuje klucz główny zbioru X i odpowiadający mu klucz obcy zbioru Y),
- w pozostałych przypadkach stosowana jest następująca formuła:

$$lr(X \triangleright\triangleleft Y) = \frac{1}{sel} \cdot \min(lr(X), lr(Y)) \tag{5}$$

Symbol *sel* reprezentuje współczynnik unikalności wartości atrybutów występujących w wyrażeniu łączącym. Definiowany jest jako iloraz liczby wartości unikalnych atrybutu przez liczebność zbioru. Dla uproszczenia przyjęto jednakową wartość współczynnika *sel* dla wszystkich zbiorów. Jest ona jednym z parametrów przeprowadzonych badań.

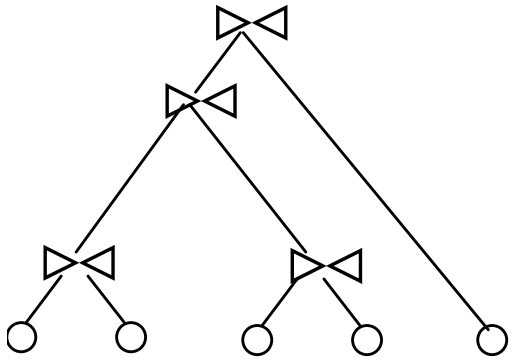
Przy oszacowanej liczebności zbioru wynikowego $lr(X \triangleright\triangleleft Y)$, znanych liczebnościach zbiorów uczestniczących w złączeniu $lr(X)$, $lr(Y)$ i mocy obliczeniowej v węzła, w którym odbywa się złączenie, koszt wykonania operacji złączenia K_Z zostanie oszacowany według następującej formuły zaproponowanej w pracy [1]:

$$K_Z = \frac{lr(X) + lr(Y) + lr(X \triangleright\triangleleft Y)}{v} \tag{6}$$

Do wyznaczenia kosztu transmisji K_T zbioru Z między dwoma węzłami sieci A i B konieczna jest znajomość liczby przesyłanych rekordów $lr(Z)$, ich długości $dr(Z)$ oraz przepustowości łącza $v_{A\to B}$. Koszt transmisji K_T zostanie oszacowany zgodnie z formułą zaczerpniętą z pracy [1]:

$$K_T = \frac{lr(Z) \cdot dr(Z)}{v_{A\to B}} \tag{7}$$

Do obliczenia wartości funkcji celu dla pojedynczego rozwiązania rozważanego problemu wykorzystano strukturę drzewiastą (rys. 4).



Rys. 4. Przykładowa struktura drzewiasta służąca do wyznaczenia wartości funkcji celu
Fig. 4. A tree structure used for evaluation of the goal function

Liście struktury zawierają charakterystykę zbiorów pierwotnych, natomiast w wierzchołkach zapisane są informacje o poszczególnych operacjach złączenia i ich rezultatach. Pojedyncza krawędź łączy liść lub wierzchołek, opisujący powstanie danego zbioru z wierzchołkiem przyporządkowanym złączeniu z udziałem tego zbioru, reprezentując przesyły zbiorów między węzłami sieci. Każdy wierzchołek jest zwieńczeniem poddrzewa obejmującego określony fragment rozwiązania i zawiera informacje pozwalające na obliczenie wartości funkcji celu dla tego fragmentu. Przez stopniowe kumulowanie wartości funkcji celu kolejne elementy struktury są wypełniane w jednym przebiegu – od liści w kierunku korzenia.

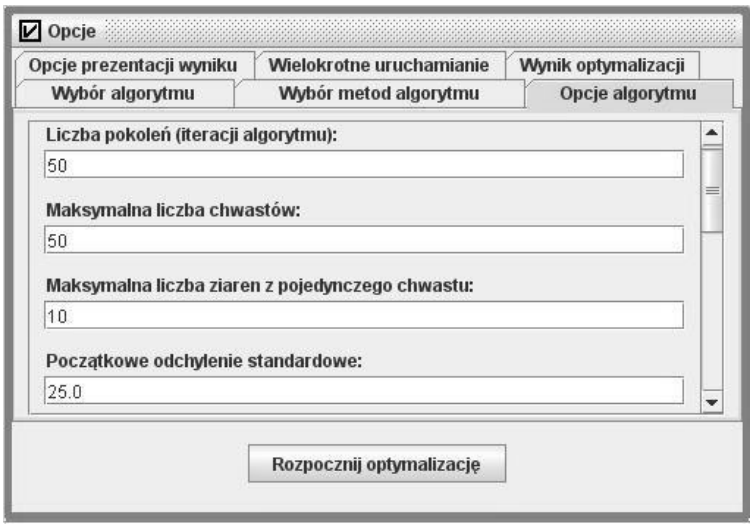
Wartość funkcji przystosowania, odgrywająca w przypadku algorytmu IWO istotną rolę w określeniu liczby nasion rozsiewanych przez daną roślinę, obliczana jest jako odwrotność funkcji celu.

4. Badania eksperymentalne

W celu przeprowadzenia eksperymentów z wykorzystaniem opracowanego algorytmu rozbudowano aplikację stanowiącą środowisko badawcze opisane w pracy [3]. Fragment okna głównego programu, prezentujący niektóre parametry istotne dla działania algorytmu IWO, przedstawiono na rys. 5.

Celem pierwszej fazy badań było wyznaczenie wartości parametrów algorytmu IWO: liczby populacji i ich liczebności, liczby nasion rozsiewanych przez pojedynczy chwast, zakresu wartości odchylenia standardowego dla odległości między chwastem macierzystym i potomnym oraz współczynnika modulacji nieliniowej, pozwalających na jak najskuteczniejszą realizację rozważanego zadania optymalizacji. Rezultaty poszczególnych prób oceniano pod kątem następujących kryteriów: czasu pracy programu, czasu potrzebnego do uzyskania najlepszego wyniku, populacji, która dała najlepszy wynik oraz odpowiadającej mu wartości funkcji celu. Wybrane na tej podstawie wartości parametrów algorytmu IWO zostały w dru-

giej fazie badań użyte w eksperymentach dających podstawy do konfrontacji z rezultatami prób wykonanych w analogicznym środowisku z użyciem algorytmu ewolucyjnego.



Rys. 5. Fragment okna głównego aplikacji
Fig. 5. A fragment of the application interface

Eksperymenty przeprowadzono zarówno dla rozproszonej, jak i dla scentralizowanej bazy danych. Zapytanie reprezentowane było przez graf złączeń o kształcie gwiazdy, łańcuchowym lub nieregularnym. Liczba zbiorów pierwotnych tworzących graf złączeń w kolejnych seriach prób przyjmowała jedną z wartości ze zbioru {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}. Zróżnicowano charakterystykę zbiorów pierwotnych (rozmiary rekordów i liczebności), węzłów (moc obliczeniowa) i fragmentów sieci (przepustowość). Dla każdego rodzaju bazy danych i grafu złączeń o zadanym kształcie, złożonego z określonej liczby zbiorów pierwotnych, przeprowadzono serię 100 prób. Jako kryterium zakończenia pojedynczej próby przyjęto przetworzenie określonej liczby populacji. Eksperymenty wykonano posługując się stacją roboczą o następujących parametrach: procesor Intel Core2 Quad Q6600 2.4GHz, pamięć RAM 2GB 800MHz. Dla grupy parametrów związanych z modelem obliczeniowym dla funkcji celu przyjęto następujące wartości liczbowe:

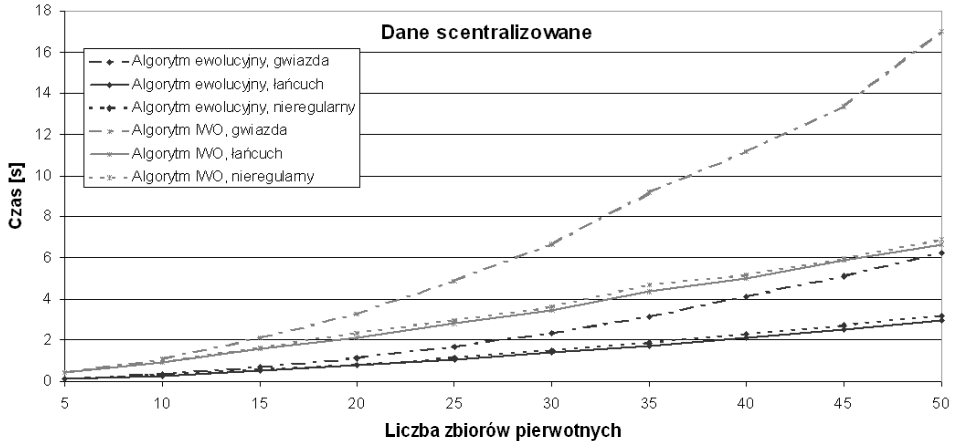
- współczynnik redukcji długości rekordu przy złączeniu $\eta = 0.9$,
- współczynnik unikalności wartości atrybutów występujących w wyrażeniu łączącym $sel = 0.8$,
- prawdopodobieństwa transformacji numeru węzła i numeru zbioru – odpowiednio: 1/3 i 2/3.

Efektem pierwszej fazy badań stały się następujące ustalenia dotyczące wartości parametrów algorytmu IWO:

- liczebność populacji roślin – stała wartość 60 w kolejnych populacjach,
- liczba populacji $iter_{max} = 60$,

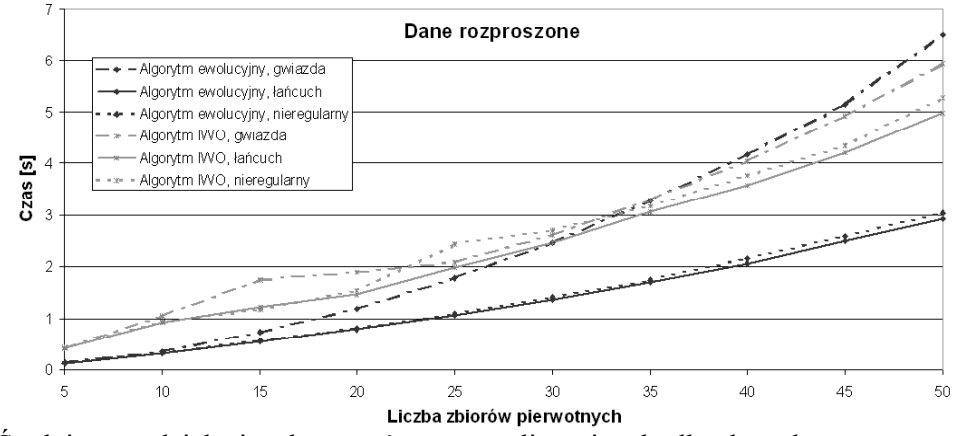
- minimalna i maksymalna liczba nasion rozsiewanych przez pojedynczy chwast – odpowiednio: 1 i 10,
- początkowa (σ_{init}) i końcowa (σ_{fin}) wartość odchylenia standardowego – odpowiednio: 5 i 1,
- współczynnik modulacji nieliniowej $m = 3$.

Wyniki uzyskane w drugiej fazie badań zestawiono z rezultatami badań wykonanych z użyciem algorytmu ewolucyjnego. Na rys. 6 i 7 zaprezentowano zależność czasu obliczeń od liczby zbiorów pierwotnych, odpowiednio dla danych scentralizowanych i rozproszonych.



Rys. 6. Średni czas działania algorytmów optymalizacyjnych dla danych scentralizowanych w zależności od liczby zbiorów pierwotnych

Fig. 6. The average running time of optimization algorithms for centralized data depending on the number of original sets



Rys. 7. Średni czas działania algorytmów optymalizacyjnych dla danych rozproszonych w zależności od liczby zbiorów pierwotnych

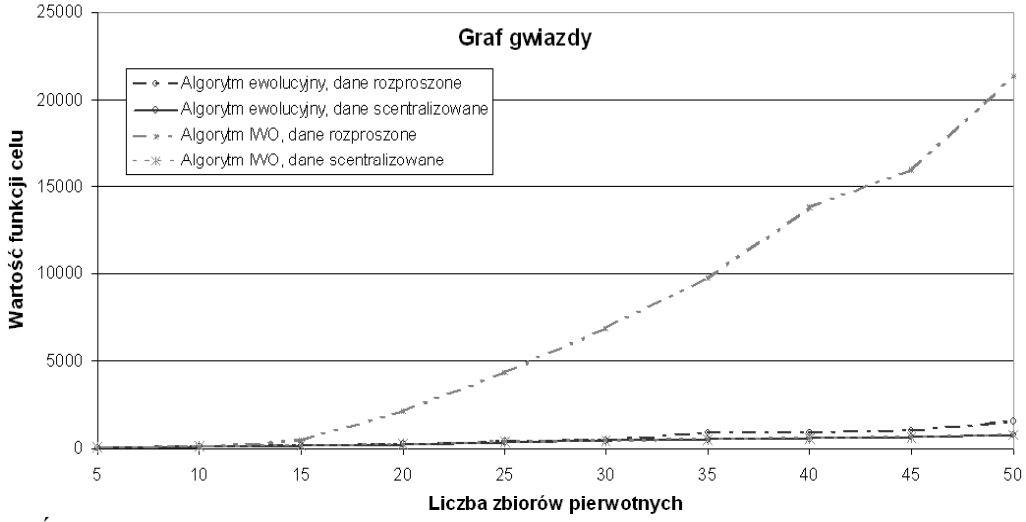
Fig. 7. The average running time of optimization algorithms for distributed data depending on the number of original sets

Z analizy rysunków 6 i 7 wynika, że zadanie wyznaczenia porządku i miejsc realizacji złączeń dla określonej liczby zbiorów pierwotnych było w zdecydowanej większości prób realizowane w krótszym czasie przez algorytm ewolucyjny. Przyczyny tego faktu należy upatrywać przede wszystkim w wykorzystaniu operatora krzyżowania. Krzyżowanie uważane

jest bowiem za operator genetyczny o zasadniczym znaczeniu [7], natomiast mutację – odpowiednik transformacji – ocenia się jako operator o znaczeniu drugorzędnym [8]. Idea algorytmu IWO wyklucza zastosowanie krzyżowania, gdyż operator ten wymaga posłużenia się dwoma osobnikami macierzystymi, podczas kiedy transformacja – tylko jednym.

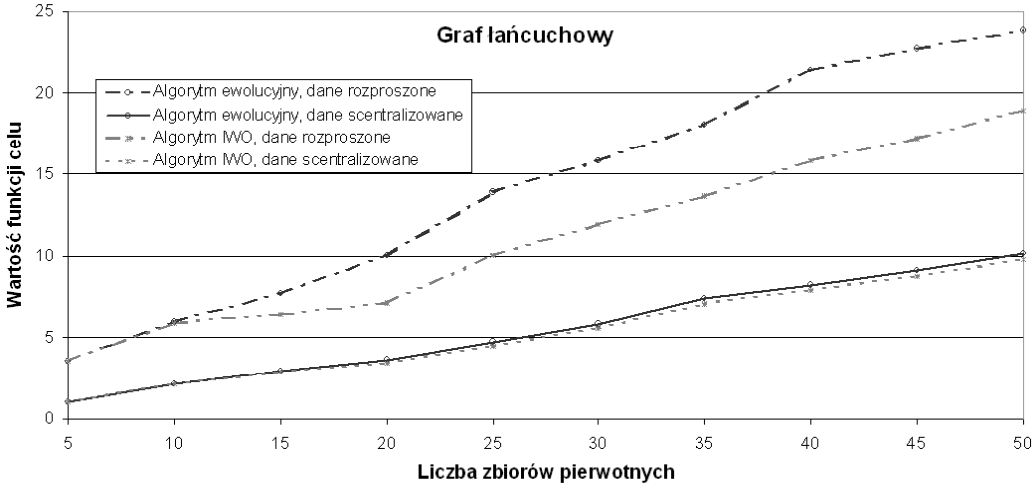
Dla pełnego porównania obu algorytmów wskazane jest dokonanie analizy jakości dawanych przez nie rozwiązań. Służyć będzie temu zestawienie rysunków prezentujących średnią wartość funkcji celu obliczoną przez określony algorytm dla danych scentralizowanych lub rozproszonych przy grafie złączeń o kształcie gwiazdy (rys. 8), łańcuchowym (rys. 9) lub nieregularnym (rys. 10). Należy zwrócić uwagę na fakt, że za rozwiązania o wyższej jakości należy uznać te, które charakteryzują się mniejszą wartością funkcji celu, rozumianej jako szacunkowy czas oczekiwania użytkownika na wynik końcowy zapytania.

Dane do wykresów przedstawionych na rysunkach 8, 9 i 10 zebrano testując oba algorytmy w konfiguracjach parametrów pozwalających na zapewnienie, aby czasy trwania pojedynczej próby dla obu algorytmów rozwiązujących to samo zadanie były jak najbardziej zbliżone do siebie.



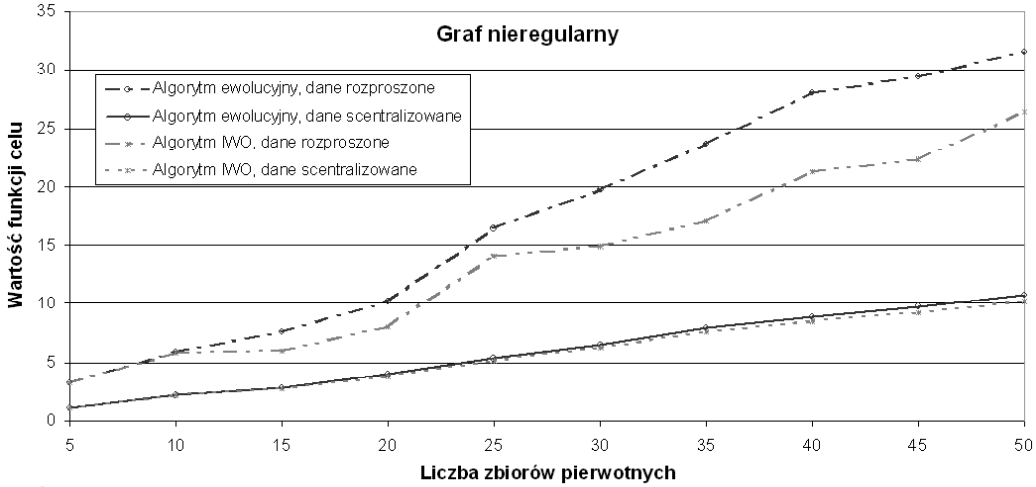
Rys. 8. Średnia wartość funkcji celu dla grafu złączeń o kształcie gwiazdy w zależności od liczby zbiorów pierwotnych

Fig. 8. The average value of the goal function for star-shaped join graph depending on the number of original sets



Rys. 9. Średnia wartość funkcji celu dla grafu złączeń o kształcie łańcuchowym w zależności od liczby zbiorów pierwotnych

Fig. 9. The average value of the goal function for chain-shaped join graphs depending on the number of original sets



Rys. 10. Średnia wartość funkcji celu dla grafu złączeń o kształcie nieregularnym w zależności od liczby zbiorów pierwotnych

Fig. 10. The average value of the goal function for irregular join graphs depending on the number of original sets

W większości przypadków rozwiązania uzyskiwane za pomocą algorytmu IWO są lepszej jakości, co szczególnie widoczne jest przy porównaniu przebiegów z rysunków 9 i 10 dla danych rozproszonych. Od tej ogólnej tendencji odbiegają rozwiązania dawane przez algorytm IWO dla grafów złączeń o kształcie gwiazdy, zwłaszcza dla danych rozproszonych. Uzasadnieniem przewagi algorytmu ewolucyjnego w tym przypadku jest sposób działania operatora krzyżowania, który zdecydowanie szybciej znajduje odpowiedni węzeł dla operacji złączenia aniżeli losowanie stosowane w algorytmie IWO. Znaczna różnica rzędu wielkości funkcji celu między rysunkami 8 a 9 i 10 wynika z użytych danych wejściowych (przede wszystkim liczebności zbiorów pierwotnych).

5. Podsumowanie

Z przeprowadzonych badań wynikają następujące wnioski:

- dłuższy czas pracy algorytmu IWO równoważony jest przez lepszą jakość generowanych przez niego rozwiązań, zwłaszcza dla zapytań reprezentowanych przez grafy złączeń o kształcie łańcuchowym lub nieregularnym,
- przewaga algorytmu ewolucyjnego uwidacznia się w przypadku zapytań operujących na danych scentralizowanych, gdzie czas pracy algorytmu ewolucyjnego jest krótszy, natomiast jakość rozwiązań generowanych przez oba algorytmy jest zbliżona, aczkolwiek w zdecydowanej większości przypadków nieco wyższa dla rozwiązań dawanych przez algorytm IWO.

BIBLIOGRAFIA

1. Chen Y.: A Systematic Method for Query Evaluation in Distributed Heterogeneous Databases. Journal of Information Science and Engineering, Vol. 16, No. 4, 2000.
2. Josiński H.: Model realizacji zapytania dla danych rozproszonych. Wysokowydajne sieci komputerowe Tom 1, Wydawnictwa Komunikacji i Łączności, Gliwice 2005.
3. Kostrzewa D., Josiński H.: Planowanie procesu scalania danych rozproszonych za pomocą algorytmu ewolucyjnego. Bazy danych. Rozwój metod i technologii. Tom 1: Architektura, metody formalne i zaawansowana analiza danych. Wydawnictwa Komunikacji i Łączności, Gliwice, 2008.
4. Lanzelotte R., S., G., Valduriez P., Zaït M.: On the Effectiveness of Optimization Search Strategies for Parallel Execution Spaces. Proceedings of the 19th VLDB Conference, Dublin 1993.
5. Mallahzadeh A., R., Oraizi H., Davoodi-Rad Z.: Application of the Invasive Weed Optimization Technique for Antenna Configurations. Progress in Electromagnetics Research, 2008.
6. Mehrabian R., Lucas C.: A novel numerical optimization algorithm inspired from weed colonization. Ecological Informatics Vol. 1, Issue 4, 2006.
7. Pawlak M.: Algorytmy ewolucyjne jako narzędzie harmonogramowania produkcji. Wydawnictwo Naukowe PWN, Warszawa 1999.
8. Rutkowski L.: Metody i techniki sztucznej inteligencji. Wydawnictwo Naukowe PWN, Warszawa 2006.
9. Sepehri Rad H., Lucas C.: A Recommender System based on Invasive Weed Optimization Algorithm. IEEE Congress on Evolutionary Computation, Singapore 2007.

10. Ullman J., D., Widom J.: Podstawowy wykład z systemów baz danych. Wydawnictwa Naukowo-Techniczne, Warszawa 1999.

Recenzent: Prof. dr hab. inż. Mieczysław Muraszkiewicz

Wpłynęło do Redakcji 3 lutego 2009 r.

Abstract

The goal of the project was to adapt the idea of the Invasive Weed Optimization (IWO) algorithm to the problem of predetermining the progress of distributed data merging process and to compare the results of the conducted experiments with analogical outcomes produced by the evolutionary algorithm. The main differences between both compared algorithms constituted by operators used for transformation of individuals and for creation of a new population were taken into consideration during the implementation of the IWO algorithm. The construction of an environment for experimental research made it possible to carry out a set of tests to explore the characteristics of the tested algorithms. The results of the conducted experiments formed the main topic of analysis.

Adresy

Daniel KOSTRZEWA: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, Daniel.Kostrzewa@polsl.pl.
Henryk JOSIŃSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, Henryk.Josinski@polsl.pl.