ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1999
Seria: INFORMATYKA z. 37 Nr kol. 1421

Piotr BAJERSKI
Politechnika Śląska, Instytut Informatyki

EFEKTYWNOŚĆ PRZETWARZANIA W JĘZYKU SQL ZAPYTAŃ PRZESTRZENNYCH WYKORZYSTUJĄCYCH APROKSYMACJE OBIEKTÓW

Streszczenie. W artykule omówiono efektywność przetwarzania w języku SQL zapytań przestrzennych wykorzystujących aproksymacje obiektów. Wykorzystano przy tym koncepcje zaprezentowane w pracy [1], dotyczące przechowywania w relacyjnej bazie danych aproksymacji obiektów przestrzennych drzewem czwórkowym uporządkowanym krzywą fraktalną N-Peano. Dla przedstawionych zapytań podano czasy ich realizacji w SZBD Access dla trzech schematów relacji oraz zaproponowano i porównano metody przyśpieszania ich wykonania.

EFFICIENCY OF PROCESSING IN SQL LANGUAGE SPATIAL QUERIES BASED ON OBJECTS APPROXIMATION

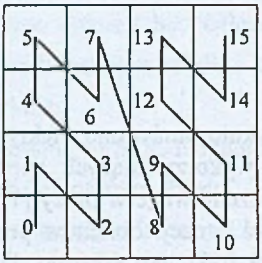
Summary. The paper presents efficiency of processing in SQL language spatial queries based on objects approximation. This work is based on ideas presented in [1], concerning storing spatial objects approximated by means of a quadtree ordered by a N–Peano fractal curve in a relational database. Execution time of presented queries in DBMS Access for three different relation schema and methods of their speeding up are given.

1. Wstęp

W artykule [1] omówiono możliwości wykorzystania języka SQL do formułowania zapytań przestrzennych operujących na aproksymacjach obiektów drzewem czwórkowym uporządkowanym krzywą N-Peano. W tym opracowaniu przedstawiono ocenę efektywności przetwarzania w DBMS Access zapytań opartych na tych ideach. W badaniach wykorzystano dane przestrzenne opisujące przebieg granic administracyjnych oraz rozkład zanieczyszczenia

powietrza na terenie województwa katowickiego. Obiekty administracyjne i rozkłady zanieczyszczenia zostały aproksymowane drzewem czwórkowym. W celu efektywnego przechowania ich w bazie danych aproksymacje zostały uporządkowane krzywą N-Peano.

Drzewo czwórkowe (ang. quadtree) jest strukturą danych klasy drzewo [8, 13, 15], powstałą poprzez rekurencyjny podział przestrzeni 2-wymiarowej na cztery części o takich samych wymiarach. Metoda ta daje najlepsze wyniki, jeżeli drzewo zbudowane jest z kwadratów o boku będącym potęgą liczby 2. Podczas aproksymacji obiektu drzewem proces podziałów kończy się, gdy cały kwadrat leży na jego obszarze lub gdy osiągnięto założoną granicę rozdzielczości. Liniowym drzewem czwórkowym będzie nazywane drzewo czwórkowe, którego elementy zostały uporządkowane krzywą fraktalną.



Rys. 1. Krzywa N-Peano
Fig. 1. N-Peano curve

Krzywa, która wykorzystuje fraktale do wypełnienia przestrzeni, jest nazywana krzywą fraktalną. Krzywą fraktalną zapewniającą proste przejście ze współrzędnych na klucz jest krzywa N-Peano [9, 10], nazywana również sekwencją Mortona, krzywą N i krzywą Z (rys. 1). Klucz N-Peano, określający położenie kwadratu na krzywej, tworzony jest poprzez przeplot bitów poszczególnych współrzędnych. W badaniach przyjęto konwencję, że parzyste bity klucza Peano odpowiadają współrzędnej X, a nieparzyste współrzędnej Y.

W eksperymentach wykorzystano dwa schematy relacji Peano omówione w [1] oraz schemat umożliwiający przechowanie kwadratów elementarnych:

- schemat 1 – minimalny klucz Peano i długość boku:

  SchemRelPeano1 (IdObiektu, KluczPeano, DługoscBoku, Atrybuty)

  gdzie:

  - *IdObiektu* - identyfikator kodowanego obiektu,
  - *KluczPeano* - klucz Peano lewego dolnego kwadratu elementarnego,
  - *DługoscBoku* - długość boku kwadratu, wyrażona liczbą kwadratów elementarnych,
  - *Atrybuty* - atrybuty opisowe, związane z danym fragmentem obiektu o identyfikatorze *IdObiektu*.

- schemat 2 – zakres kluczy Peano:

  SchemRelPeano2 (Id_obiektu, KluczPeanoPoczatku, KluczPeanoKonca, Atrybuty)

  gdzie:
  - *KluczPeanoPoczatku* - klucz Peano lewego dolnego kwadratu elementarnego (minimalna wartość klucza Peano odpowiadająca kwadratowi),
  - *KluczPeanoKonca* - klucz Peano prawego górnego kwadratu elementarnego (maksymalna wartość klucza Peano odpowiadająca kwadratowi),
  - pozostałe atrybuty mają takie samo znaczenie jak w schemacie 1.
- schemat 3 – kwadraty elementarne:

  Schemat3 (Id_obiektu, KluczPeano, Atrybuty)

  gdzie:
  - *KluczPeano* - klucz Peano kwadratu elementarnego,
  - pozostałe atrybuty mają takie samo znaczenie jak w schemacie 1.

Każda krotka relacji o przedstawionych schematach opisuje jeden kwadrat tworzący aproksymację.

Relacje o schemacie 3, w których w kolejnych krotkach zapamiętywane są wszystkie kwadraty elementarne tworzące aproksymację obiektów, są normalnymi relacjami w rozumieniu Codda [16], ponieważ nie można takiej krotki rozbić na równoważny jej zbiór krotek.

Dla relacji Peano wyróżnia się trzy następujące poziomy poprawności (szczegółowo zostały omówione w [1]):

- pierwszy – wszystkie kwadraty są prawidłowo położone, co oznacza, że mogły powstać przy rekurencyjnym podziale przestrzeni podczas tworzenia drzewa czwórkowego,
- drugi - brak nakładających się kwadratów,
- trzeci - zwarte kwadraty, co oznacza, że kwadraty tworzące większy kwadrat są łączone.

2. Charakterystyka wykorzystywanych danych

Do badań wykorzystano aproksymacje rozkładów stężenia pyłu zawieszonego i ołowiu w powietrzu w roku 1989. Rozkłady zostały wyznaczone na podstawie wartości zmierzonych w sieci pomiarowej Wojewódzkiej Stacji Sanitarno-Epidemiologicznej w Katowicach. Aproksymacje rozkładów wygenerowano dla rozdzielczości 512 i 1024 piksele dwoma metodami wybranymi spośród dziesięciu metod przedstawionych w [2]. Pierwsza metoda, oznaczana przez *M3*, zapewniająca kompromis między dokładnością i szybkością, wykorzystuje interpolację lokalną. Decyzja o zakwalifikowaniu kwadratu do przedziału bądź jego dalszym podziale jest podejmowana na podstawie wartości wyznaczonych dla środka kwadratu i środ-

ków czterech mniejszych kwadratów. Druga metoda, oznaczana przez *M10*, daje dokładniejszy obraz rozkładu, ale jest dużo bardziej złożona obliczeniowo i daje większą liczbę kwadratów w aproksymacji. Tworzy ona aproksymację na podstawie rozkładu wygenerowanego metodą krigingu [5, 7].

Dla obydwu zanieczyszczeń utworzono po cztery przedziały, których granice są wielokrotnościami dopuszczalnego stężenia zanieczyszczenia. Tabele 1, 2 i 3 przedstawiają liczbę kwadratów występujących w aproksymacjach poszczególnych rozkładów i typów obiektów administracyjnych. Aproksymacja obiektów administracyjnych obejmowała 1 województwo, 84 gminy i 13 największych miast.

Tabela 1

Liczby kwadratów przypadających na poszczególne przedziały w aproksymacjach rozkładu pyłu i ołowiu metodą *M3* i *M10* dla rozdzielczości 512 i 1024 i schematów 1 i 2

Przedział	Pyl M3		Pyl M10		Pb M3		Pb M10	
	512	1024	512	1024	512	1024	512	1024
0	697	1149	1633	3501	896	1445	2122	4669
1	2124	3281	3775	7767	2169	3323	4617	9633
2	1083	1826	1533	3410	1182	1986	2055	4547
3	65	117	83	205	327	524	469	1050
Suma	3969	6373	7024	14883	4574	7278	9263	19899

Tabela 2

Liczba kwadratów przypadających na poszczególne przedziały w aproksymacjach rozkładu pyłu i ołowiu metodą *M3* i *M10* dla rozdzielczości 512 i 1024 i schematu 3

Przedział	Pyl M3		Pyl M10		Pb M3		Pb M10	
	512	1024	512	1024	512	1024	512	1024
0	11633	46308	12588	50358	13198	52530	14919	59682
1	73319	292269	72012	286991	68319	272334	65431	260745
2	6174	24689	6541	25977	6998	27954	8173	32497
3	221	876	206	816	2832	11324	2824	11218
Suma	91347	364142	91347	364142	91347	364142	91347	364142

Aby uprościć zapis zapytań i zwiększyć efektywność ich przetwarzania, aproksymacje rozkładów są przechowywane w osobnych tablicach. Poniżej przedstawiono schematy tablic dla trzech wyróżnionych schematów.

- schemat 1:

  AproksObAdmS1 (IdObAdm, KP, DlBk)

  AproksRozklZanS1 (KP, DlBk, NrPrzedz)

- schemat 2:

  AproksObAdmS2 (IdObAdm, KPpocz, KPkon)

  AproksRozklZanS2 (KPpocz, KPkon, NrPrzedz)

- schemat 3:

  AproksObAdmS3 (*IdObAdm, KP*)

  AproksRozklZanS3 (*KP, NrPrzedz*)

gdzie *NrPrzedz* określa numer przedziału, do którego został zakwalifikowany dany kwadrat.

Dane opisowe obiektów administracyjnych są przechowywane w tablicy:

DaneAdm (IdObAdm, NazwaObAdm, TypObAdm)

Tabela 3

Histogramy liczby kwadratów przypadających na poszczególne typy obiektów administracyjnych dla rozdzielczości 512 i 1024 i trzech schematów

Typ obiektu administracyjnego	Reprezentacja 1 i 2		Reprezentacja 3	
	512	1024	512	1024
Województwo	2259	4688	91347	364142
Gminy	14212	30088	96496	374359
Miasta	699	1535	2007	7385
Suma	17170	36311	189850	745886

3. Zapytania

Korzystając z przedstawionych w [1] propozycji zapisu zapytań sprawdzających wzajemne relacje topologiczne aproksymacji obiektów, można w zapytaniu w języku SQL wyrazić dowolne warunki (jeżeli nie czyni się rozróżnienia między wnętrzem obszaru a jego brzegiem) na wzajemne relacje topologiczne między dwoma rozkładami oraz rozkładami i innymi obiektami przestrzennymi. Rozkład przestrzenny dany za pomocą aproksymacji obszarów należących do poszczególnych wyróżnionych przedziałów można traktować jako zbiór aproksymacji obiektów przestrzennych, tak że obiektowi odpowiada obszar należący do danego przedziału. Obiekt taki może składać się z wielu rozłącznych części i zawierać dziury. Poszczególne obiekty muszą być rozłączne – dany kwadrat elementarny może należeć tylko do jednego przedziału. Sytuacja taka nie zachodzi dla obszarów administracyjnych, gdzie po aproksymacji wektorów tworzących granice administracyjne kwadraty elementarne odpowiadające granicy były dołączane do graniczących obiektów. Poniżej podano zapis kilku wybranych zapytań dotyczących relacji między rozkładem zanieczyszczenia a obiektami administracyjnymi, które wydają się ważne z praktycznego punktu widzenia.

Zapytanie 1

Podaj gminy, na terenie których wartość zanieczyszczenia należała do przedziału 3. Warunki zapytania spełniają relacje topologiczne od 2 do 8. Po eliminacji z zapytań warunków odpowiadających za rozróżnienie tych przypadków powstaje następujące zapytanie:

```
SELECT da.IdObAdm, da.NazwaObAdm
FROM DaneAdm da
WHERE da.TypObAdm = 'GMINA' AND EXISTS ( SELECT *
        FROM AproksObAdm ap, AproksRozklZan r
        WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
                da.IdObAdm = ap.IdObAdm AND r.Przedzial = 3);
```

Zapytanie to można zapisać bez zapytania zagnieżdżonego:

```
SELECT DISTINCT da.IdObAdm, da.NazwaObAdm
FROM DaneAdm da, AproksObAdm ap, AproksRozklZan r
WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
        da.IdObAdm = ap.IdObAdm AND r.Przedzial = 3 AND da.TypObAdm = 'GMINA';
```

Zapytanie 2

Podaj gminy, na terenie których wartość zanieczyszczenia należała do przedziału 2 lub wyższego. Zapytanie to wygląda tak samo jak poprzednie – zmienia się tylko warunek na numer przedziału.

Zapytanie 3

Podaj gminy, na terenie których nie występowało zanieczyszczenie z przedziału 3. Warunki spełniają obiekty pozostające w relacji topologicznej 1.

```
SELECT da.IdObAdm, da.NazwaObAdm
FROM DaneAdm da
WHERE da.TypObAdm = 'GMINA' AND NOT EXISTS (
        SELECT *
        FROM AproksObAdm ap, AproksRozklZan r
        WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
                da.IdObAdm = ap.IdObAdm AND r.Przedzial = 3);
```

Zapytanie 4

Podaj gminy, na terenie których poziom zanieczyszczenia należy tylko do przedziału 0. Jest to pytanie o zachodzenie relacji 5 lub 6. Pytanie to można przeformułować i zapytać się o wszystkie gminy, które są rozłączne (relacja topologiczna nr 1) ze wszystkimi przedziałami poza przedziałem nr 0.

```
SELECT da.IdObAdm, da.NazwaObAdm
FROM DaneAdm da
WHERE da.TypObAdm = 'GMINA' AND NOT EXISTS (
        SELECT *
        FROM AproksObAdm ap, AproksRozklZan r
        WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
                da.IdObAdm = ap.IdObAdm AND r.Przedzial > 0);
```

Zapytanie 5

Podaj gminy, na terenie których wartość zanieczyszczenia należała do przedziału 2 lub wyższego. Dla każdej znalezionej gminy podaj wielkość obszaru gminy należącego do tych przedziałów (wyrażonego liczbą elementarnych kwadratów).

```
SELECT da.IdObAdm, da.NazwaObAdm, SUM (
        r.KPkon * SGN (1 – SGN (r.KPkon – ap.KPkon)) +
        ap.KPkon * (1- SGN (1 + SGN (ap.KPkon - r.KPkon))) -
        r.KPpocz * SGN (1 + SGN (r.KPpocz – ap.KPpocz)) -
        ap.KPpocz * (1 – SGN (1 – SGN (ap.KPpocz – r.KPpocz))) + 1) AS PowZan
FROM DaneAdm da, AproksObAdm ap, AproksRozklZan r
WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
        da.IdObAdm = ap.IdObAdm AND r.Przedzial >= 2 AND da.TypObAdm = 'GMINA'
GROUP BY da.IdObAdm, da.NazwaObAdm;
```

Zapytanie 6

Podaj gminy, których przynajmniej 30% obszaru było zanieczyszczone na poziomie należącym do przedziału 2 lub wyższego. Dla każdej znalezionej gminy podaj procent obszaru gminy należącego do tych przedziałów.

```
SELECT da.IdObAdm, da.NazwaObAdm, SUM (
    r.KPkon * SGN (1 – SGN (r.KPkon – ap.KPkon)) +
    ap.KPkon * (1- SGN (1 + SGN (ap.KPkon - r.KPkon))) -
    r.KPpocz * SGN (1 + SGN (r.KPpocz – ap.KPpocz)) -
    ap.KPpocz * (1 – SGN (1 – SGN (ap.KPpocz – r.KPpocz)))
    + 1) / wo.Obszar * 100 AS PowZan
FROM DaneAdm da, AproksObAdm ap, AproksRozklZan r,
        (SELECT ao.IdOb, SUM KPkon –KPpocz + 1) AS Obszar
         FROM AproksObAdm ao, DaneAdm da
        WHERE da.IdObAdm = ao.IdObAdm AND lTypObiektu = 'GMINA'
        GROUP BY ao.IdOb) wo
WHERE ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz AND
        da.IdObAdm = ap.IdObAdm AND da.TypObAdm = 'GMINA' AND
        r.Przedzial >= 2 AND wo.IdObAdm = ap.IdObAdm
GROUP BY da.IdObAdm, da.NazwaObAdm, wo.Obszar
HAVING SUM (
    r.KPkon * SGN (1 – SGN (r.KPkon – ap.KPkon)) +
    ap.KPkon * (1- SGN (1 + SGN (ap.KPkon - r.KPkon))) -
    r.KPpocz * SGN (1 + SGN (r.KPpocz – ap.KPpocz)) -
    ap.KPpocz * (1 – SGN (1 – SGN (ap.KPpocz – r.KPpocz)))
    + 1) > 0.3 * wo.Obszar;
```

4. Wyniki

Tabele 4, 5 i 6 przedstawiają czas wykonania każdego z sześciu badanych zapytań dla rozkładu średniorocznego stężenia pyłu i ołowiu w rozbiciu na rozdzielczości i metody generacji aproksymacji rozkładów. Tabela 4 przedstawia wyniki pomiarów dla danych przechowywanych w relacjach o schemacie 1, tablica 5 przechowywanych w relacjach o schemacie 2, a tablica 6 przechowywanych w relacjach o schemacie 3. Eksperymenty były prowadzone na komputerze Pentium II z procesorem 266 MHz i pamięcią operacyjną 64MB.

Pomiar czasu był wykonywany z dokładnością do jednej sekundy, tak więc wyniki dla zapytań wykonujących się szybko są obarczone dużym błędem. Można przyjąć, że wyniki poniżej 3 s oznaczają, że zapytania są wykonywane w trybie interakcyjnym.

Dla wszystkich tablic występujących w zapytaniach założone zostały indeksy złożone na kluczach oraz atrybutach, po których występowały złączenia. Przeprowadzono eksperymenty z różnymi konfiguracjami indeksów, ale brak było widocznego wpływu organizacji indeksów na czas wykonywania zapytań. Wykorzystanie frazy *EXISTS* w zapytaniu wydłużało czas jego wykonania w stosunku do wersji bez zapytania zagnieżdżonego. Dlatego, gdy było to możliwe, stosowano wersje bez pytań zagnieżdżonych.

Tabela 4

Czas wykonania zapytań, w sekundach, dla schematu 1, w zależności od użytych danych

Pytanie	Rozdzielczość 512				Rozdzielczość 1024			
	Pył M3	Pył M10	Pb M3	Pb M10	Pył M3	Pył M10	Pb M3	Pb M10
1	18	22	79	113	66	115	265	534
2	294	421	370	615	1086	1967	1138	2949
3	17	23	71	104	63	113	234	476
4	451	802	462	871	1486	3494	1529	3820
5	299	426	372	615	1067	1977	1328	2886
6	296	425	373	622	1075	1997	1328	2928

Tabela 5

Czas wykonania zapytań, w sekundach, dla schematu 2, w zależności od użytych danych

Pytanie	Rozdzielczość 512				Rozdzielczość 1024			
	Pył M3	Pył M10	Pb M3	Pb M10	Pył M3	Pył M10	Pb M3	Pb M10
1	9	11	41	59	34	60	139	277
2	154	228	195	331	553	1033	685	1513
3	9	13	38	55	33	56	113	229
4	217	407	233	439	761	1817	768	1971
5	156	216	196	317	561	1016	673	1486
6	160	233	200	341	566	1049	701	1547

Tabela 6

Czas wykonania zapytań, w sekundach, dla schematu 3, w zależności od użytych danych

Pytanie	Rozdzielczość 512				Rozdzielczość 1024			
	Pył M3	Pył M10	Pb M3	Pb M10	Pył M3	Pył M10	Pb M3	Pb M10
1	2	2	2	2	5	5	7	7
2	9	8	9	9	38	37	36	38
3	2	2	3	3	5	4	15	15
4	28	26	25	25	97	93	96	91
5	9	8	9	8	37	36	38	38
6	10	8	9	9	35	32	33	33

Aproksymacje rozkładów generowane metodami *M3* i *M10* znacząco różnią się liczbą kwadratów. Jednakże wyniki tych samych zapytań dla danych generowanych różnymi metodami różniły się najwyżej o jedną gminę. Dodanie do zapytania ograniczenia, że warunki nałożone na przedziały wartości mają dotyczyć przynajmniej zadanego procenta powierzchni gminy, pozwala wyeliminować te różnice.

5. Poprawa efektywności przetwarzania zapytań

Jakkolwiek przedstawione trzy schematy pozwalają przechować te same informacje, to jednak różnią się pod względem czasu przetwarzania i ilości miejsca potrzebnego do przechowania danych. Przy przyśpieszaniu wykonania zapytań skoncentrowano się na zapytaniach operujących na tablicach o schemacie 2. Z jednej strony taki format zapisu danych zajmuje znacznie mniej miejsca niż wykorzystanie schematu 3, z drugiej strony zapytania te są szybsze od zapytań operujących na tablicach o schemacie 1.

Zaproponowane metody przyśpieszania wykonania zapytań można podzielić na dwie grupy:

- wykorzystujące konstrukcje języka SQL, należą do niej partycje i agregaty,
- wymagające rozszerzenia języka SQL o fragmenty kodu w języku proceduralnym, należą do nich złączenie liniowe i statystyki (do których obliczenia wymagane jest użycie pętli).

5.1. Złączenie relacji Peano o złożoności liniowej

Przy przetwarzaniu zapytań, w których występują warunki sprawdzające nakładanie się obiektów zapisanych w relacjach Peano (spełniających przynajmniej drugi poziom poprawności), serwer relacyjnej bazy danych nie jest w stanie wykorzystać szczególnych własności tych

danych. Wyznaczenie części wspólnej aproksymacji dwóch obiektów jest równoważne wykonaniu złączenia tablic zawierających te aproksymacje z warunkiem nakładania się zakresów kluczy.

Jeżeli nie istnieje indeks, to serwer musi sprawdzić wszystkie kombinacje kwadratów i liczba porównań jest równa iloczynowi liczby kwadratów w obydwu aproksymacjach. Jeżeli istnieje indeks na kluczu Peano, to serwer może go wykorzystać w algorytmie przeglądania zagnieżdżonego. Zewnętrzna pętla przechodzi po wszystkich elementach aproksymacji obiektu pierwszego, wewnętrzna przebiega elementy aproksymacji obiektu drugiego, takie że ich minimalny klucz Peano jest mniejszy lub równy maksymalnemu kluczowi Peano bieżącego elementu aproksymacji obiektu pierwszego. Liczba porównań średnio będzie równa połowie iloczynu liczby elementów obydwu aproksymacji i dalej jest kwadratowa.

Wykorzystując własność, że kwadraty tworzące aproksymacje obiektu nie nakładają się, można zaproponować bardziej efektywne rozwiązanie o liniowej złożoności. W zewnętrznej pętli przy przejściu do następnego elementu aproksymacji obiektu pierwszego nie trzeba cofać się do pierwszego elementu drugiej aproksymacji, lecz można pozostać na elemencie bieżącym. Dodatkowo można pomijać elementy, które nie mają szans na złączenie.

W przypadku badania, czy aproksymacje nakładają się, wykonywanie złączenia można przerwać po znalezieniu pierwszego elementu części wspólnej.

Poniższy fragment programu w pseudo-Visual Basic'u znajduje gminy, na terenie których zanieczyszczenie spełnia zadany warunek (dla uproszczenia usunięto kod związany z obsługą sytuacji błędnych). Zbiór rekordów *rstAproksRozkl* zawiera kwadraty tworzące aproksymację rozkładu spełniające zadany warunek na zanieczyszczenie. Są one uporządkowane według klucza Peano. Zbiór rekordów *rstAproksObAdm* zawiera aproksymacje obiektów administracyjnych. Dla każdego obiektu kwadraty tworzące jego aproksymacje są uporządkowane według klucza Peano. Zbiór rekordów *rstGminy* zawiera wszystkie obiekty administracyjne typu gmina.

```
Do Until rstGminy.EOF
    ' znajdź pierwszy (o najmniejszym kluczu Peano) kwadrat aproksymacji bieżącej gminy
    rstAproksObAdm.Seek ">=", rstGminy!IIdObAdm
    rstAproksRozkl.MoveFirst                ' wróć na pierwszy element aproksymacji rozkładu
    ' przejdź po wszystkich elementach aproksymacji danego ob. adm.
    Do While Not rstAproksObAdm.EOF And Not rstAproksRozkl.EOF And
            rstAproksObAdm!IdObAdm = rstGminy!IIdObAdm
        If rstAproksObAdm!KPpocz <= rstAproksRozkl!KPkon And
            rstAproksObAdm!KPkon >= rstAproksRozkl!KPpocz Then
            'znaleziono część wspólną – wyświetl informacje o gminie
            Exit Do         ' zakończ sprawdzanie dla bieżącego obiektu
        End If
        ' pomiń elementy aproksymacji, które nie mają szans na złączenie
```

```
            Do While Not rstAproksRozkl.EOF And
                rstAproksRozkl!KPkon < rstAproksObAdm!KPpocz
                rstAproksRozkl.MoveNext
            Loop
            If rstAproksRozkl.EOF Then Exit Do End If
            Do While Not rstAproksObAdm.EOF AND rstAproksObAdm!IdObAdm =
                rstGminy!IIdObAdm And rstAproksObAdm!KPkon < rstAproksRozkl!KPpocz
                rstAproksObAdm.MoveNext
            Loop
            End If
        Loop
        rstGminy.MoveNext
    Loop
```

Tabela 7

Czas wykonania zapytań wykorzystujących liniowe złączenie dla schematu 2, w zależności od użytych danych

Pytanie	Rozdzielczość 512				Rozdzielczość 1024			
	Pył M3	Pył M10	Pb M3	Pb M10	Pył M3	Pył M10	Pb M3	Pb M10
1	1	1	1	1	1	2	2	2
2	2	2	1	3	3	5	3	6
4	3	5	3	6	4	12	5	14

5.2. Wykorzystanie agregatów

Inne podejście do przyśpieszania wykonania zapytań o obszary, na których zanieczyszczenie daną substancją należy do pewnych przedziałów, opiera się na idei wcześniejszego wyliczania pewnych wartości (nazywanych agregatami), z których można skorzystać w trakcie realizacji zapytania, eliminując czasochłonne obliczenia. Rozważając dane wykorzystywane w przykładach, można wcześniej wyliczyć dla każdego obiektu administracyjnego, jaki jego obszar jest zajmowany przez poszczególne przedziały zanieczyszczenia. Takie agregaty należy oczywiście tworzyć niezależnie dla każdego zanieczyszczenia i dla każdej rozdzielczości. Przedstawione agregaty mogą być składowane w tablicy o schemacie:

TabAgregObAdmRozklZan (IdObAdm, NrPrzedz, Wielkosc)

gdzie:

- *Wielkosc* – zawiera liczbę kwadratów elementarnych należących do obiektu o identyfikatorze *IdOb*, które zostały zakwalifikowane do przedziału *NrPrzedz*,
- pozostałe atrybuty mają takie samo znaczenie jak wcześniej.

W przypadku gdy dany przedział nie leży na terenie danego obiektu administracyjnego, można zapamiętać wartość 0 lub nie wprowadzać wiersza do tablicy. W badaniach wykorzystano drugie rozwiązanie.

Przedstawione zapytania można automatycznie przetłumaczyć na postać korzystającą z tablicy agregatów. Zapytanie 2 przyjęłoby następującą postać:

SELECT DISTINCT da.IdObAdm, da.NazwaObAdm
FROM DaneAdm AS da, TabAgregObAdmRozklZan AS a
WHERE da.IdObAdm = a. IdObAdm AND a.NrPrzedz >= 2 AND da.TypObAdm = 'GMINA';

Wszystkie zapytania przy użyciu agregatów wykonywały się poniżej 1 s. Obliczanie takich agregatów można zapisać w języku SQL. Poważną wadą tego podejścia jest długi czas obliczania agregatów - dla rozkładu pyłu generowanego metodą *M3* przy rozdzielczości *512* wynosił on 636 s, a dla rozkładu ołowiu generowanego metodą *M10* przy rozdzielczości *1024* wynosił 6563 s. Tak długi czas eliminuje możliwość zadawania zapytań wykorzystujących agregaty zaraz po wygenerowaniu rozkładu. Z drugiej strony, jeżeli zapytania nie są zadawane zaraz po jego generacji, to agregaty można przygotować wcześniej, w dogodnej chwili.

5.3. Wykorzystanie statystyk

Czas wykonania przedstawionych zapytań zależy od liczby kwadratów, których nakładanie należy sprawdzić. Wyeliminowanie części obiektów z zapytania pozwoliłoby na znaczące zredukowanie jego złożoności obliczeniowej, ponieważ zostałyby wyeliminowane wszystkie kwadraty tworzące ich aproksymacje. Prostym (obliczeniowo) testem, czy dwa obiekty mogą się nakładać, jest sprawdzenie, czy odpowiadające im zakresy kluczy Peano nakładają się. Zakres klucza jest wyznaczany przez minimalny i maksymalny klucz dla wszystkich kwadratów aproksymujących dany obiekt. Wadą takiego podejścia jest możliwość występowania dużych luk w wartościach klucza, gdy obiekt leży na obszarze kwadratów o dużych różnicach wartości klucza. Bardziej efektywne jest podejście dopuszczające tworzenie kilku zakresów klucza dla danego obiektu, tak że każdy z nich obejmuje kwadraty o bliskich wartościach klucza.

Poniżej podano przykład wykorzystania statystyk w zapytaniu o gminy, na terenie których występuje zanieczyszczenie na poziomie przedziału 3. Zapytanie przetwarzane jest dwuetapowo. Na pierwszym etapie w pomocniczej tablicy *TabIdObAdmPom* gromadzone są identyfikatory gmin, na obszarze których może wystąpić zanieczyszczenie z przedziału 3. Przy wykonaniu tego zapytania wykorzystywane są statystyki. Na drugim etapie, dla wybranych na pierwszym etapie gmin, wykorzystując aproksymacje, sprawdza się, czy rzeczywiście występuje zanieczyszczenie z przedziału 3.

```
' znalezienie identyfikatorów gmin, na obszarze których może
' wystąpić zanieczyszczenie z przedziału 3
INSERT INTO TabIdObAdmPom
SELECT DISTINCT IdObAdm
FROM StatObAdm so, StatRozklZan sr, DaneAdm da
WHERE so.Ob = da.IdOb AND da.TypObiektu = 'GMINA' AND
        so.KPpocz <= sr.KPkon AND so.KPkon >= sr.KPpocz AND sr.Przedzial = 3;
' wykonanie zapytania z wykorzystaniem wcześniej znalezionych identyfikatorów gmin
SELECT DISTINCT da.IdOb, da.tNazwaOb
FROM DaneAdm da, TabIdObAdmPom pi, AproksObAdm ap, AproksRozklZan r
WHERE pi.IdOb = da.IdOb AND da.IdOb = ap.IdOb AND r.Przedzial = 3 AND
        ap.KPpocz <= r.KPkon AND ap.KPkon >= r.KPpocz;
```

Eksperymenty z zapytaniami wykorzystującymi statystyki zostały przeprowadzone dla rozkładu ołowiu utworzonego metodą *M10* dla rozdzielczości *512*. Statystyki były generowane poza bazą danych. Poszczególne elementy statystyki dotyczące obiektu były tak tworzone, że odległość między nimi na krzywej N-Peano musiała przekraczać zadaną wartość graniczną. Czas wykonywania pierwszego zapytania zawierał się między 20 a 25 s dla wartości granicznej wahającej się od 50 do 1000 kwadratów. Dla drugiego zapytania czas ten wahał się od 215 s do 228 s dla takiego samego przedziału wartości granicznej.

5.4. Wykorzystanie partycji

Innym podejściem do zmniejszenia czasu wykonania zapytań jest podział danych pomiędzy kilka tablic. Przy realizacji zapytań dotyczących relacji Peano dane w naturalny sposób grupują się według położenia w przestrzeni (zakresów klucza Peano). Podział danych według tego klucza prowadzi do redukcji liczby kombinacji, jaką należy sprawdzić przy wyznaczaniu odpowiedzi. Zapytania wykorzystujące partycje wyglądają analogicznie do podanych, z tym że każde z zapytań operuje tylko na podanych partycjach. Zwracane przez nie wyniki są łączone operatorem *UNION*. Przy tworzeniu partycji tablicy przechowującej elementy relacji Peano może wystąpić konieczność podziału niektórych elementów, jeżeli leżą one na obszarze, który po podziale będzie należał do kilku tablic.

Tabela 8

Wyniki podziału aproksymacji rozkładu Pb M10 dla rozdzielczości 512 na cztery partycje

Nr partycji	Liczba elementów aproksymacji obiektów administracyjnych	Liczba elementów aproksymacji rozkładu Pb M10 512	Czas wykonania pytania 1 [s]	Czas wykonania pytania 2 [s]
1	8747	4114	9	52
2	1420	989	0	2
3	5307	3098	11	48
4	1956	1062	0	4

Eksperymenty były prowadzone na aproksymacji rozkładu ołowiu wygenerowanej metodą *M10* w rozdzielczości *512* pikseli. Aby zapewnić poprawność wykonania zapytań, dane były dzielone zgodnie z budową drzewa czwórkowego. Po podziale na cztery partycje czas realizacji zapytania 1 spadł do 21 s, a zapytania 2 do 105 s.

Po podziale dwóch najliczniejszych partycji czas realizacji zapytania 1 spadł do 10 s, a pytania 2 do 51 s. Jakkolwiek kolejny podział wyrównał różnice pomiędzy partycjami, to jednak dalej były one znaczne i czas realizacji zapytania w głównej mierze był uwarunkowany czasem realizacji zapytań składowych na dwóch partycjach.

6. Podsumowanie

Relacje o schematach 1 i 2 zajmują tyle samo miejsca, jednakże zapytania do danych przechowywanych w tablicach o schemacie 2 są prawie dwukrotnie szybciej wykonywane.

W zapytaniach wykorzystujących postać 3 badanie wzajemnego położenia obiektów nie wymaga warunków na zakres wartości i serwer jest w stanie w pełni wykorzystać dostępne mechanizmy dostępu do danych. Zapytania oparte na schemacie 3 są znacznie szybsze od zapytań opartych na schemacie 2 (od 4.5 do 47 razy). Jednakże okupione to zostało dużą liczbą rekordów potrzebnych do przechowania danych. Dla aproksymacji obiektów administracyjnych zastosowanie schematu 3 powodowało 11-krotny wzrost liczby rekordów dla rozdzielczości *512* i 20-krotny dla rozdzielczości *1024*. Dla aproksymacji rozkładów zanieczyszczenia maksymalny wzrost wyniósł 57.

Przy dwukrotnym wzroście rozdzielczości dla schematów 1 i 2 liczba kwadratów wzrasta około dwukrotnie (wzrost liniowy lub mniejszy - dla metody *M3* wzrost o 1.6). Dla schematu 3 wzrost ten jest czterokrotny (kwadratowy), przy czym wielkość rekordu dla schematu 3 jest tylko nieznacznie mniejsza niż dla pozostałych dwóch schematów.

Czas wykonania zapytań dla reprezentacji 2 jest proporcjonalny do iloczynu liczby kwadratów aproksymacji obiektów administracyjnych i rozkładów. Zależność ta jest widoczna zarówno przy zmianie metody generacji rozkładu, jak i przy zmianie rozdzielczości. Przy zmianie rozdzielczości z *512* na *1024* czas wykonania zapytań dla schematu 2 średnio wzrastał 4-krotnie, a dla schematu 3 3,8-krotnie.

Największe przyśpieszenie wykonania zapytań daje zastosowanie złączenia liniowego. W zależności od danych dochodziło ono do 150 razy. Wykonanie zapytań przy użyciu tego algorytmu na tablicy o schemacie 2 daje kilkakrotne przyśpieszenie w stosunku do zapytań operujących na danych przechowywanych w tablicach o schemacie 3. Ważną jego cechą jest prawie liniowy wzrost czasu wykonania zapytań przy kwadratowym wzroście rozpatrywanego obszaru.

W podejściu wykorzystującym partycje zapytanie składa się z wielu zapytań połączonych operatorem *UNION*. Zapytania składowe działają na niezależnych tablicach i dopiero ich wyniki są łączone. Głównym problemem przy tworzeniu partycji jest równoważenie ich zapełnienia (tabela 8). Dla przedstawionych danych jest to trudne, ponieważ województwo ma nieregularny kształt, tak że ponad połowa powierzchni kwadratu, do którego zostało wpisane, leży poza obszarem województwa. Dlatego przy przyjętym algorytmie tworzenia partycji zgodnie z podziałem drzewa czwórkowego tablice odpowiadające centrum województwa są silnie wypełnione, a tablice odpowiadające granicom zawierają mało kwadratów. Drugim czynnikiem jest zmienność rozkładu zanieczyszczenia. W rejonach o dużej zmienności powstaje znacznie więcej kwadratów niż w rejonach o małej zmienności.

Wykorzystanie zaproponowanych statystyk na kluczach Peano daje tym większe przyśpieszenie, im mniejsza liczba obiektów spełnia kryteria zapytania i im więcej obiektów jest eliminowanych na pierwszym etapie wykonania zapytania. Dla przedstawionych pomiarów przyśpieszenie wykonania zapytania 1 wyniosło 3, a zapytania 2 – 1,54. Spośród 84 gmin warunki pytania 1 spełnia 20 gmin, na pierwszym etapie zostały wyeliminowane 62 gminy. Dla zapytania 2 warunki spełnia 41 gmin, na pierwszym etapie zostało wyeliminowanych 38 gmin.

Warto podkreślić, że zaproponowane złączenie liniowe może być połączone z wykorzystaniem przedstawionych statystyk i partycjami. W takim przypadku statystyki byłyby tworzone dla każdej z partycji. Na pierwszym etapie dla każdego zapytania składowego byłyby wyznaczane obiekty, które mogą spełniać warunki zapytania. Na drugim etapie korzystając z aproksymacji wybranych obiektów byłyby wyznaczane odpowiedzi częściowe, które po scaleniu stanowiłyby odpowiedź na zapytanie. Każda z zaprezentowanych metod przyśpieszałaby wykonanie zapytania.

Wykorzystanie partycji jest w naturalny sposób predestynowane do przetwarzania równoległego, gdzie każde z zapytań składowych może być wykonywane przez niezależny procesor. W celu przyśpieszenia pobierania danych tablice mogą być przechowywane na niezależnych dyskach.

Podsumowując, wykorzystanie schematu 2 pozwala na efektywne przechowanie aproksymacji obiektów przestrzennych w relacyjnej bazie danych i daje duże możliwości przyśpieszania wykonania zapytań dotyczących relacji przestrzennych między tymi obiektami.

LITERATURA

1. Bajerski P.: Możliwości zapisu w języku SQL zapytań przestrzennych wykorzystujących aproksymacje obiektów. Zeszyty Naukowe Politechniki Śląskiej 1999, seria Informatyka zeszyt 37.
2. Bajerski P.: Sprawozdanie z realizacji projektu badawczego Komitetu Badań Naukowych nr 8T11C 028 12. Wykorzystanie drzew czwórkowych i algebry Peano do prezentacji i przetwarzania rozkładów zanieczyszczenia powietrza, Politechnika Śląska, Gliwice 1998.
3. Brinhhoff T., Kriegel H., Seeger B.: Efficient Processing of Spatial Joins Using R-trees. ACM SIGMOD, 1993.
4. Bruns H., Egenhofer M.: Similarity of spatial scenes. Advances in GIS Research, Taylor & Francis 1997.
5. Cressie N. A. C., Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics 1995.
6. Egenhofer M., Franzoza R.: Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems 5(2) 1991.
7. Isaaks E.H., Srivastava R.M.: An Introduction to Applied Geostatistics, Oxford University Press 1989.
8. Jankowski M.: Elementy grafiki komputerowej. WNT, Warszawa 1990.
9. Laurini R., Francoise M.: Spatial database queries: relational algebra versus computational geometry. Proceedings of the Fourth International Conference on Statistical and Scientific Database Management, Rome, Italy 1988, M. Rafamelli et al.,(eds) Berlin;Germeny: Springer Verlag. pp. 291-313.
10. Laurini R., Thompson D.: Understanding GIS, Academic Press Limited, third printing 1994.
11. Oosterom P., Vijlbrief T.: The Spatial Location Code in Advances in GIS Research II, edited by Kraak M.J. and Molenaar M., Taylor&Francis Ltd. 1997.
12. Oracle7 Spatial Data Option™ Overiew: Oracle April 1996.
13. Pavlidis T.: Grafika i przetwarzanie obrazów, WNT, Warszawa 1987.
14. Rozenshtein D., Abramovich A., Birger E.: Speeding Up SQL Queries – Characteristically, Database Programming & Design vol.8 no.10 1995.
15. Sammet H.: The Design and Analysis of Spatial Data Structures. Addison-Wesley, Reding 1989.
16. Ullman J.: Systemy baz danych. WNT, Warszawa 1988.

Recenzent: Prof. dr hab. inż. Alicja Wakulicz-Deja

Wpłynęło do Redakcji 17 grudnia 1998 r.

Abstract

The paper presents study of efficiency of processing in SQL language spatial queries based on objects approximated by means of a quadtree ordered by a N–Peano fractal curve [1]. The experiments were carried out using air pollution measurements over Upper Silesia. Two-dimensional approximations of distribution of some pollutants were created and stored in a database. Two methods of quadtree construction, differing in accuracy and number of tree elements [2], were used.

Three methods of storing spatial objects approximated by means of a quadtree ordered by N-Peano fractal curve in a relational database were examined and compared with respect to query time execution and disk space consumption. The experiments were carried out on an MS Access database.

Relational server is unable to use special properties of the approximation during spatial join. Therefore an algorithm with linear complexity was proposed. Experiments showed that it gave dramatic query execution speedup. However its usage demands extending queries in SQL with code in a procedural language. Three other methods of query execution speeding up, which do not require procedural support, were developed and evaluated: data partitioning according to Peano key, statistics on Peano keys, and aggregates.