

Standard SQL/MM: SQL Multimedia and Application Packages

Krzysztof Jankiewicz, Marek Wojciechowski

Politechnika Poznańska, Instytut Informatyki
Krzysztof.Jankiewicz@cs.put.poznan.pl
Marek.Wojciechowski@cs.put.poznan.pl

Streszczenie

SQL/MM jest nowym standardem uzupełniającym język SQL o obsługę zaawansowanych typów danych. Celem referatu jest przedstawienie głównych idei standardu SQL/MM, a w szczególności jego części poświęconych przetwarzaniu danych tekstowych, przestrzennych i obrazów w bazach danych. Przeanalizowana zostanie również zgodność rozwiązań oferowanych w tym zakresie przez Oracle10g ze standardem SQL/MM.

1. Wprowadzenie

Od współczesnych systemów zarządzania bazami danych wymaga się już nie tylko efektywnej obsługi danych alfanumerycznych (łańcuchów znaków, wartości liczbowych, dat), ale również danych multimedialnych, charakteryzujących się złożoną strukturą i dużym rozmiarem. W praktyce, mianem danych multimedialnych określa się obecnie najczęściej zakodowane w postaci zrozumiałej dla komputera obrazy, sekwencje audio, filmy wideo (sekwencje wideo, zazwyczaj uzupełnione o sekwencję audio i opcjonalnie napisy w postaci tekstowej) oraz heterogeniczne dokumenty multimedialne (np. prezentacje). Niekiedy do danych multimedialnych zalicza się również duże dokumenty tekstowe i dane przestrzenne (np. geograficzne).

Ponieważ standard języka SQL nie zawiera konstrukcji do obsługi takich danych jak multimedia, duże obiekty tekstowe, czy też dane przestrzenne, środowiska zajmujące się tworzeniem oprogramowania do przetwarzania tych specjalistycznych typów danych rozpoczęły pracę nad propozycjami rozszerzenia języka SQL o potrzebne im elementy. Organizacje odpowiedzialne za standard SQL zwróciły jednak uwagę na fakt, że ewentualne rozszerzenia SQL dedykowane dla poszczególnych rodzajów danych mogą być niekompatybilne ze sobą. Najłatwiejszym do zauważenia potencjalnym konfliktem był konflikt słów kluczowych. Przykładowo, słowo kluczowe *CONTAINS*, używane jest zarówno w kontekście tekstowych baz danych (do wskazania, że dane słowo lub fraza zawiera się w danym fragmencie tekstu), jak i baz przestrzennych i multimedialnych (do wskazania, że jeden obiekt zawiera w sobie inny obiekt).

Ze względu na wspomniane wyżej problemy zarzucono koncepcję rozszerzania języka SQL w zakresie wsparcia dla baz danych tekstowych, przestrzennych i multimedialnych. Zwyciężyła koncepcja opracowania nowego standardu, obejmującego specyfikacje bibliotek opartych o typy obiektowe SQL99 [7], przeznaczonych do obsługi poszczególnych specjalistycznych rodzajów danych i aplikacji. Nowy standard natychmiast stał się znany pod nazwą SQL/MM [6] („MM” od „MultiMedia”). Dzięki oparciu specyfikacji SQL/MM na obiektowych typach SQL, funkcjonalność bibliotek jest w sposób naturalny dostępna z poziomu poleceń języka SQL, np. poprzez wywołania metod bibliotecznych typów obiektowych w wyrażeniach języka SQL.

Celem niniejszego artykułu jest dokonanie przeglądu standardu SQL/MM ze szczególnym uwzględnieniem części dotyczących danych przestrzennych, tekstowych i obrazów oraz analiza zgodności rozwiązań oferowanych w tym zakresie przez Oracle10g ze standardem. Układ artykułu przedstawia się następująco. Rozdział 2 przedstawia strukturę standardu SQL/MM i krótko opisuje jego części. Rozdziały 3, 4 i 5 poświęcone są odpowiednio specyfikacjom SQL/MM dla danych

przestrzennych, tekstowych i obrazów. Rozdział 6 przedstawia funkcjonalność Oracle10g w zakresie przetwarzania zaawansowanych rodzajów danych na tle standardu SQL/MM. Rozdział 7 stanowi krótkie podsumowanie.

2. Struktura standardu SQL/MM

Podobnie jak SQL, SQL/MM składa się z wielu części, przy czym części SQL/MM są ze sobą raczej luźno związane w porównaniu ze specyfikacjami SQL. Pełna nazwa standardu SQL/MM brzmi: SQL Multimedia and Application Packages, co oznacza że jego części niekoniecznie muszą dotyczyć przetwarzania szeroko pojętych danych multimedialnych, ale również innych zastosowań. W chwili obecnej standard obejmuje pięć części:

- Część pierwsza – *Framework* [1] ma charakter ogólny i nie dotyczy żadnego ze specjalistycznych rodzajów danych. Zawiera ona definicje i koncepcje wspólne dla pozostałych, specjalistycznych części standardu. Część pierwsza dotyczy między innymi sposobu, w jaki inne części standardu SQL/MM wykorzystują mechanizm obiektowych typów SQL.
- Część druga – *Full-Text* [2] definiuje typy danych do składowania dużych obiektów tekstowych, skupiając się na wsparciu dla operacji wyszukiwania zadanych wzorców w tekście.
- Część trzecia – *Spatial* [3] definiuje typy danych wspierające aplikacje przetwarzające dane przestrzenne. W obecnej wersji standardu wsparcie ograniczone jest do obiektów 0-wymiarowych (punktów), 1-wymiarowych (linii) i 2-wymiarowych (figur płaskich).
- Część piąta – *Still Image* [4] definiuje typy danych do obsługi obrazów w bazie danych, ze szczególnym naciskiem na wyszukiwanie obrazów ze względu na zawartość.
- Część szósta – *Data mining* [5], w przeciwieństwie do części drugiej, trzeciej i piątej, nie opisuje złożonych rodzajów danych. Jest ona jedyną w obecnej postaci standardu częścią dotyczącą standardowych pakietów aplikacyjnych i definiuje typy danych do realizacji zadań eksploracji danych.

Należy zwrócić uwagę, że wymienione wyżej części standardu, poświęcone poszczególnym dziedzinom zastosowań, nie zostały zatwierdzone jednocześnie. Najwcześniej, bo już w 1999 roku wykrystalizowała się specyfikacja dotycząca danych przestrzennych, w 2000 roku – specyfikacja dotycząca danych tekstowych; następnie w 2001 roku została zatwierdzona specyfikacja dla obrazów i wreszcie, jako najmłodsza obecnie część standardu, w 2002 roku pojawiła się specyfikacja poświęcona eksploracji danych. Specyfikacje standardu SQL/MM ulegają ciągłym ewolucjom i z pewnością nie mają charakteru ostatecznego. Aktualne wersje części dotyczących danych przestrzennych, tekstowych i obrazów pochodzą z 2003 roku.

Nie ma w standardzie SQL/MM części czwartej. Miała ona dotyczyć ogólnych operacji matematycznych (*General Purpose Facilities*), ale prace nad nią zarzucono kilka lat temu. Być może w przyszłości pojawią się natomiast kolejne części SQL/MM dotyczące multimediiów, np. danych wideo. Ewentualne poszerzenie SQL/MM o nowe części będzie zależało od zapotrzebowania i odzewu środowisk bazodanowych na istniejące specyfikacje SQL/MM.

3. Dane przestrzenne (Spatial)

SQL/MM *Spatial* definiuje obiektowe typy danych i ich metody do przetwarzania danych przestrzennych tj. dotyczących geometrii, lokalizacji i topologii obiektów. SQL/MM *Spatial* jest związany z dwoma innymi, rozwijanymi równolegle, standardami dla danych przestrzennych opracowywanymi przez ISO Technical Committee TC 211 i Open GIS Consortium.

SQL/MM *Spatial* obecnie wspiera dane 0-wymiarowe (punkty), 1-wymiarowe (linie) i 2-wymiarowe (figury płaskie). W rzeczywistych zastosowaniach pojawiają się oczywiście również obiekty 3-wymiarowe, niestety obecna specyfikacja SQL/MM ich nie uwzględnia. Dużą wagę specyfikacja przykładu do przestrzennych układów odniesienia. Większość z nich opisuje geografie naszej planety i jej regionów, co jest konsekwencją faktu, że typowymi użytkownikami wymagającymi przetwarzania danych przestrzennych są władze państwowe i lokalne oraz duże

korporacje, operujące na danych geograficznych. Specyfikacja SQL/MM *Spatial* jest szczególnie zorientowana na przetwarzanie danych w systemach informacji geograficznej (GIS), ale obszar jej zastosowań jest szerszy i obejmuje również np. projektowanie układów elektronicznych.

SQL/MM *Spatial* specyfikuje zbiór typów danych, tworzących kilka hierarchii. Podstawową hierarchię typów tworzą typy reprezentujące poszczególne figury (kształty) geometryczne. Korzeniem tej hierarchii jest abstrakcyjny typ *ST_Geometry*. Inne typy abstrakcyjne, pełniące rolę węzłów pośrednich w hierarchii to *ST_Curve*, *ST_Surface*, *ST_MultiCurve* i *ST_MultiSurface*. W hierarchii występują ponadto następujące nieabstrakcyjne podtypy: *ST_Point*, *ST_LineString*, *ST_CircularString*, *ST_CompoundCurve*, *ST_CurvePolygon*, *ST_Polygon*, *ST_GeomCollection*, *ST_MultiPoint*, *ST_MultiLineString* i *ST_MultiPolygon*.

Spośród pozostałych typów SQL/MM *Spatial* należy wymienić *ST_SpatialRefSystem* służący do opisu przestrzennych układów odniesienia oraz typy *ST_Angle* i *ST_Direction* reprezentujące odpowiednio kąty i kierunki.

Metody typów SQL/MM *Spatial* można podzielić na 4 grupy [13]:

- Metody do konwersji z/do zewnętrznych formatów danych (tekstowych, binarnych i formatu GML – Geography Markup Language), np. *ST_MPointFromGML*, *ST_LineFromGML*, *ST_AsGML*;
- Metody do odczytu właściwości i miar dla obiektów geometrycznych, np. *ST_Boundary*, *ST_Length*, *ST_Area*;
- Metody do porównywania obiektów geometrycznych: *ST_Equals*, *ST_Disjoint*, *ST_Intersects*, *ST_Crosses*, *ST_Overlaps*, *ST_Touches*, *ST_Within*, *ST_Contains* i *ST_Distance*;
- Metody do tworzenia nowych obiektów geometrycznych, np. *ST_Difference*, *ST_Intersection* i *ST_Union*.

W celu zilustrowania sposobu korzystania z typów danych SQL/MM *Spatial*, rozważmy poniższy schemat tabeli do składowania informacji o miastach:

```
CREATE TABLE miasta
(nazwa VARCHAR(40),
 lokalizacja ST_Polygon);
```

Poniższe zapytanie zwraca pole powierzchni obszaru, który zajmuje Warszawa. Do wyznaczenia pola powierzchni wielokąta służy metoda *ST_Area*, zdefiniowana w typie *ST_Surface*, z którego pośrednio dziedziczy wykorzystany w tabeli *MIASTA* typ *ST_Polygon*.

```
SELECT lokalizacja.ST_Area
FROM miasta
WHERE nazwa = 'Warszawa';
```

4. Dane tekstowe (Full-Text)

SQL/MM *Full-Text* dotyczy przetwarzania dokumentów tekstowych, które od tradycyjnych łańcuchów znaków odróżnia przede wszystkim znacznie większy rozmiar, ale również obecność struktury (podział na zdania i akapity) oraz charakter operacji wyszukiwania. O ile dla prostych łańcuchów znaków wyszukiwanie sprowadza się do prostego dopasowywania do wzorca, w przypadku danych tekstowych kryteria wyszukiwania dotyczą obecności słów kluczowych lub fraz i uwzględniają odmianę, wymowę, a nawet znaczenie poszczególnych słów czy fraz w kontekście konkretnego języka. Mechanizmy przeszukiwania tekstów objęte standardem SQL/MM są szczególnie odpowiednie dla języków, które łatwo poddają się analizie komputerowej w zakresie wyodrębniania poszczególnych słów i zdań. Do tej klasy języków należą języki zachodnie (w tym np. angielski, niemiecki, polski), w przypadku których słowa oddzielone są białymi spacjami, a zdania znakami interpunkcyjnymi. (Inaczej jest np. w języku japońskim, gdzie wyodrębnienie słów z tekstu wymaga analizy kontekstu.)

SQL/MM *Full-Text* definiuje kilka typów obiektowych, z których podstawowe znaczenie ma typ *FullText*, służący do reprezentacji dokumentów tekstowych. Spośród metod typu *FullText* największe znaczenie mają: *Contains* umożliwiająca zaawansowane dopasowywanie tekstu do wzorca i *Score*, która dla danego dokumentu i wzorca zwraca miarę zgodności (ang. *relevance*) dokumentu ze wzorcem.

W celu ilustracji możliwości oferowanych przez typ *FullText* rozważmy poniższą tabelę służącą do składowania dokumentów tekstowych:

```
CREATE TABLE dokumenty
(nr INTEGER,
 tresp FULLTEXT);
```

Poniższe zapytanie zwraca dokumenty, w których występuje słowo 'standard':

```
SELECT nr
FROM dokumenty
WHERE tresp.Contains(' "standard" ') = 1;
```

Metoda *Contains* akceptuje wiele klas wzorców, bardziej zaawansowanych niż sprowadzających się do wyszukania jednego słowa. Poniżej przedstawiono niektóre z nich:

- Wystąpienie konkretnej frazy, np. `tresp.Contains(' "standard języka" ')`
- Wzorce ze znakami '_' i '%', np. `tresp.Contains(' "standard%" ')`
- Wzorce dopuszczające synonimy, np. `tresp.Contains(THESAURUS "computer science" EXPAND SYNONYM TERM OF "list" ')`
- Wystąpienie co najmniej jednego słowa kluczowego lub jednej frazy z listy, np. `tresp.Contains(' "standard", "język%" ')`
- Wystąpienie formy danego słowa/frazy wg reguł odmiany danego języka (domyślny jest angielski), np. `tresp.Contains('STEMMED FORM OF POLISH "standard języka"')`
- Wystąpienie słów/fraz w danym kontekście, np. `tresp.Contains(' "standard" IN SAME SENTENCE AS "SQL" ')`
- Wzorce złożone za pomocą operatorów logicznych AND, OR i NOT, np. `tresp.Contains(' "SQL/MM" & ("Text" | "Spatial") & NOT "Image" ')`
- Wzorce odwołujące się do tematyki tekstu, np. `tresp.Contains(' IS ABOUT "standard języka zapytań" ')`
- Wzorce odwołujące się do brzmienia w danym języku (domyślny jest angielski), np. `tresp.Contains(' SOUNDS LIKE "sequel" ')`

Metoda *Contains* służy do przeprowadzania binarnych testów zgodności z wzorcem, nie informując o tym, na ile dobrze dany wzorec opisuje konkretny dokument. Jeśli celem przeszukania kolekcji dokumentów jest dokonanie ich rankingu w świetle zadanych kryteriów wyszukiwania, należy skorzystać z metody *Score*. Metoda ta zwraca nieujemną liczbę zmiennoprzecinkową, przy czym większa wartość oznacza lepsze dopasowanie. Metoda *Score* przyjmuje takie same parametry co *Contains*, inne jest znaczenie i zakres zwracanych przez nią wartości, np. `tresp.Score(' "standard", "język", "SQL" ')`.

Dla metod *Contains* i *Score* typu *FullText*, standard przewiduje istnienie ich odpowiedników w postaci funkcji SQL o tych samych nazwach, których pierwszym argumentem jest dokument, a drugim wzorec.

5. Obrazy (Still Image)

Specyfikacja SQL/MM *Still Image* definiuje strukturalne typy SQL umożliwiające składowanie obrazów w bazie danych, ich różnego rodzaju modyfikacje oraz wyszukiwanie obrazów spełniających kryteria selekcji odwołujące się do atrybutów wizualnych obrazu. Dodatkowo, dla niektórych metod tych strukturalnych typów, standard przewiduje istnienie ich odpowiedników w postaci funkcji SQL.

W standardzie SQL/MM obrazy są reprezentowane za pomocą typu *SI_StillImage*. *SI_StillImage* składa się z dwuwymiarowych obrazów w postaci kolekcji pikseli. Standard zakłada, że obrazy mogą być składane w różnych formatach (JPEG, GIF, TIFF, itd.), zależnie od możliwości konkretnej implementacji. Oprócz samego obrazu (składanego w postaci dużego obiektu binarnego – BLOB), *SI_StillImage* zawiera również takie informacje o obrazie jak format i wymiary w pikselach. Metody typu *SI_StillImage* pozwalają m.in. na skalowanie i obroty obrazu, obcinanie obrazu, konwersję formatów oraz generację miniaturki obrazu w mniejszej rozdzielczości (ang. thumbnail).

Oprócz podstawowego typu *SI_StillImage*, SQL/MM w zakresie obsługi obrazów definiuje również kilka typów służących do reprezentacji różnych właściwości obrazu. Typ *SI_AverageColor* reprezentuje średni kolor obrazu, *SI_ColorHistogram* dostarcza informacji o udziale kolorów w obrazie, *SI_PositionalColor* reprezentuje lokalizację poszczególnych kolorów na obrazie, a *SI_Texture* służy do zapamiętania informacji o tzw. teksturze obrazu. Dla każdego z wymienionych typów istnieje metoda typu *SI_StillImage* zwracająca daną właściwość wskazanego obrazu. Ponadto, typ *SI_StillImage* i każdy z typów reprezentujących właściwości obrazu posiada metodę *SI_Score*, służącą do wyznaczenia odległości między dwoma obrazami w kontekście danej właściwości. Odległość przyjmuje wartości większe lub równe 0, większa odległość oznacza większą różnicę.

W celu zilustrowania sposobu korzystania z typów danych *StillImage*, rozważmy tabelę *WZORY* do składowania kolekcji obrazów:

```
CREATE TABLE wzory
(nazwa VARCHAR2(40),
obraz SI_StillImage);
```

Poniższe przykładowe zapytanie wykorzystujące typy SQL/MM zwraca wzory podobne w sensie tekstury do zadanego wzorca. Wykorzystana w klauzuli *WHERE* funkcja *SI_findTexture* wyznacza teksturę nowego wzoru w postaci wartości typu *SI_Texture*. Na rzecz uzyskanej w ten sposób reprezentacji tekstury podanego obrazu, wołana jest następnie metoda *SI_Score*, która wyznacza odległość pod względem tekstury między podanym obrazem a obrazami składowanymi w bazie danych w kolumnie typu *SI_StillImage*.

```
SELECT nazwa
FROM wzory
WHERE SI_findTexture((SELECT obraz FROM wzory WHERE nazwa = 'nowy')).
SI_Score(obraz) < 1.2
```

Poszczególne właściwości obrazu (w postaci obiektów *SI_AverageColor*, *SI_ColorHistogram*, *SI_PositionalColor*, *SI_Texture*) mogą być wyznaczone online gdy zachodzi konieczność porównania obrazu z innym obrazem względem danego kryterium, co ilustruje powyższy przykład. Alternatywą jest składowanie wyznaczonych właściwości obrazów w bazie danych w kolumnach typów *SI_AverageColor*, *SI_ColorHistogram*, *SI_PositionalColor*, *SI_Texture*.

Oprócz wspomnianych wyżej typów danych, z których każdy umożliwia testy podobieństwa względem jednej właściwości, specyfikacja SQL/MM *Still Image* na potrzeby wyszukiwania podobnych obrazów definiuje jeszcze typ *SI_FeatureList*. Typ ten umożliwia reprezentację zbioru właściwości obrazu z przypisaniem wag do poszczególnych właściwości. Kolejne właściwości dodawane są do obiektu *SI_FeatureList* metodami *SI_SetFeature*, których pierwszym parametrem jest właściwość (w postaci obiektu *SI_AverageColor*, *SI_ColorHistogram*, *SI_PositionalColor*, *SI_Texture*), a drugim podana w postaci liczby z przedziału <0.0, 1.0> waga. *SI_FeatureList* zawiera metodę *SI_Score*, która dla danego obrazu zwraca miarę podobieństwa w postaci średniej ważonej miar podobieństwa dla właściwości uwzględnionych w danym obiekcie *SI_FeatureList*.

6. Obsługa danych przestrzennych, tekstowych i obrazów w Oracle 10g w świetle standardu SQL/MM

Oceniając możliwości systemu zarządzania bazą danych Oracle w zakresie przetwarzania zaawansowanych rodzajów danych w świetle standardu SQL/MM, należy rozpocząć od obrazów,

gdyż właśnie specyfikacja *SQL/MM Still Image* jest jedyną, którą Oracle10g implementuje. Analizując mechanizmy obsługi danych tekstowych i przestrzennych w Oracle10g, można jedynie dokonać porównania funkcjonalności oferowanej w tym zakresie przez Oracle i standard *SQL/MM*.

Implementacja *SQL/MM Still Image* w Oracle10g opiera się na pierwszej edycji standardu z 2001 roku (ISO/IEC 13249-5:2001). Obecnie dostępna jest już druga edycja standardu, z roku 2003, niemniej jednak niektóre nowe elementy w niej zawarte, zostały uwzględnione w Oracle10g jako rozszerzenia obowiązującej w trakcie prac nad Oracle10g specyfikacji *SQL/MM Still Image*.

6.1. Oracle interMedia – obsługa obrazów i innych rodzajów multimedialnych

W ramach serwera bazy danych Oracle10g za przetwarzanie danych multimedialnych (obrazów, sekwencji audio i wideo oraz heterogenicznych dokumentów) odpowiada właściwość (ang. feature) występująca pod nazwą *interMedia* [9][10], dostępna w standardowej instalacji serwera (Standard Edition). Dla wszystkich wymienionych typów multimedialnych Oracle *interMedia* w wersji 10g umożliwia składowanie binarnej zawartości w bazie danych i na zewnątrz (w systemie plików serwera lub z dostępem przez URL) oraz składowanie, przetwarzanie i automatyczną ekstrakcję metadanych o formacie (typ MIME, format danych, metoda kompresji, itp.) i zawartości merytorycznej (np. tytuł utworu i wykonawca dla danych audio). Ponadto, tylko dla obrazów, *interMedia* umożliwia proste operacje na zawartości takie jak: konwersja formatu i/lub zmiana metody kompresji, kopiowanie, skalowanie i obcinanie obrazu oraz wyszukiwanie ze względu na zawartość (ang. content-based image retrieval).

Oracle *interMedia* opiera się na obiektowo-relacyjnych własnościach serwera Oracle, wykorzystując wsparcie dla składowania w bazie danych dużych obiektów binarnych (BLOB) i możliwość definiowania obiektowych (strukturalnych) typów danych SQL. Funkcjonalność Oracle *interMedia* jest dostępna głównie poprzez 4 typy obiektowe (zdefiniowane w schemacie *ORDSYS*):

- *ORDAudio* – do składowania i przetwarzania obiektów audio;
- *ORDVideo* – do składowania i przetwarzania obiektów wideo;
- *ORDImage* – do składowania i przetwarzania obrazów;
- *ORDDoc* – do składowania heterogenicznych obiektów multimedialnych.

Typy danych *ORDAudio*, *ORDVideo*, *ORDImage*, i *ORDDoc* dostępne były już w wersjach serwera bazy danych Oracle wcześniejszych niż 10g. Nowością w Oracle10g są typy danych *SQL/MM Still Image*, umożliwiające zgodne ze standardem operacje na obrazach w bazie danych, stanowiąc alternatywę dla *ORDImage*. Implementując *SQL/MM Still Image*, Oracle oferuje pewne rozszerzenia funkcjonalności typów definiowanych przez standard, np. pola w typie *SI_StillImage* do zapamiętania właściwości wizualnych obrazu.

Mimo wsparcia dla typów *SQL/MM Still Image*, podstawowym typem danych do składowania i przetwarzania obrazów w Oracle10g jest nadal typ *ORDImage*. Oracle *interMedia* za pośrednictwem typu *ORDImage* uwzględnia przy wyszukiwaniu obrazów ze względu na zawartość następujące właściwości:

- kolor – udział poszczególnych kolorów;
- tekstura – „niskopoziomowe” wzory na obrazie, szczególnie te powtarzające się;
- kształt – kształty występujące na obrazie, rozumiane jako regiony o jednym kolorze;
- lokalizacja – reprezentująca położenie kolorów, tekstur i kształtów na obrazie (ma sens tylko w połączeniu z co najmniej jedną z pozostałych właściwości).

Właściwości wizualne obrazów składowanych jako *ORDImage* reprezentowane w postaci sygnatur (obiektów typu *ORDImageSignature*). Wszelkie operacje związane z oceną podobieństwa obrazów są wykonywane na sygnaturach, a nie bezpośrednio na obrazach. Sygnatury muszą być jawnie wyznaczone (za pomocą metody *generateSignature* typu *ORDImageSignature*). Sygnatury obrazów mogą być składowane w tabelach tak samo jak obrazy (*ORDImage*) i inne dane.

Przy wyszukiwaniu ze względu na zawartość, użytkownik przypisuje wagi poszczególnym właściwościom obrazów. Miarą podobieństwa jest ważona suma odległości między obrazami w kontekście poszczególnych właściwości. Wynik jest liczbą zmiennoprzecinkową z zakresu od 0 (dla obrazów identycznych) do 100 (dla obrazów całkowicie różnych).

Do przeprowadzania testów podobieństwa na sygnaturach służy operator *IMGSimilar*, który oprócz dwóch sygnatur otrzymuje na wejście łańcuch znaków z przypisaniem wag do właściwości

wizualnych i próg maksymalnej dopuszczalnej odległości. Operator *IMGSimilar* może być wykorzystywany w połączeniu z podrzędnym operatorem *IMGScore*, który zwraca wartość funkcji odległości wyznaczoną podczas testu podobieństwa. Dla poprawy efektywności operacji wyszukiwania obrazów podobnych do danego, Oracle InterMedia umożliwia tworzenie specjalnego typu indeksu (*ORDSYS.ORDIMAGEINDEX*) na kolumnach typu *ORDImageSignature*.

Ponieważ obecna wersja serwera Oracle udostępnia dwie rodziny typów do reprezentacji i przetwarzania obrazów, narzuca się pytanie, których z nich należy używać. Z pewnością typy *SQL/MM Still Image* potencjalnie umożliwią w przyszłości przenoszenie aplikacji na serwery baz danych innych producentów. Z kolei *ORDImage* jest bardziej efektywny, a ponadto oferuje funkcjonalność niedostępną w standardzie *SQL/MM*, np. wyszukiwanie podobieństw ze względu na kształt i możliwość zakładania indeksów na sygnaturach reprezentujących właściwości wizualne obrazów.

6.2. Oracle Text – obsługa danych tekstowych

Oracle Text [12] jest właściwością systemu zarządzania bazą danych Oracle10g wspierającą przetwarzanie danych tekstowych, dostępną zarówno w wersji Standard jak i Enterprise. Pod nazwą Oracle Text kryje się technologia umożliwiająca budowę aplikacji wyszukujących i/lub klasyfikujących dokumenty tekstowe na podstawie ich zawartości.

W odróżnieniu od standardu *SQL/MM*, Oracle Text nie wprowadza, żadnych wyspecjalizowanych typów pozwalających na przechowywanie dużych dokumentów tekstowych. Wynika to z przyjętej koncepcji dokumentu. W przypadku Oracle Text dokument niekoniecznie musi być pojedynczą kolumną tekstową o dużej objętości. Może nim być zbiór kolumn w jednej tabeli, zbiór wierszy w tabeli podrzędnej dla tabeli głównej będącej korzeniem dokumentu, może nim być plik w systemie operacyjnym lub strona dostępna poprzez sieć WWW. Kolejną różnicą jest to, że system zarządzania bazą danych Oracle nie wymaga, aby wyszukiwaniu podlegały tylko i wyłącznie dokumenty tekstowe. Umożliwia on przeszukiwanie w kontekście zawartości tekstowej również dokumentów binarnych. Dokumenty mogą posiadać najróżniejsze formaty: procesorów tekstów, arkuszy kalkulacyjnych, formatów plików bazodanowych, prezentacji, a nawet formatów graficznych. Jest to podejście znacznie wykraczające poza standard *SQL/MM* i znacznie bardziej odpowiadające rzeczywistym potrzebom użytkowników.

Kluczem do właściwego i efektywnego przeszukiwania zawartości tekstowej są indeksy. Sposób utworzenia indeksu, jego typ, użyte parametry (preferencje) determinują możliwości późniejszych zapytań. Dla przykładu parametr *Datastore* definiuje sposób przechowywania dokumentów (pojedyncza kolumna, tabela zagnieżdżona, sieć WWW, system plików itp.), a parametr *Filter* odpowiada za konwersję dokumentów do ich postaci tekstowej, w celu indeksacji jej zawartości.

Oracle dla różnych sposobów wykorzystywania zasobów tekstowych udostępnia różne specjalizowane typy indeksów. Jeden z nich o nazwie *CONTEXT* jest typem wspierającym operator *CONTAINS*, mający podobne znaczenie i funkcjonalność jak metoda *Contains* w standardzie *SQL/MM*. Zakres zastosowania operatora *CONTAINS* to głównie: wyszukiwanie konkretnej frazy, wyszukiwanie frazy przy określeniu kolejności występowania słów, wyszukiwanie niedokładne (fuzzy), wyszukiwanie wg brzmienia słów (soundex), wyszukiwanie oparte na rdzeniach słów, wyszukiwanie wykorzystujące wzorce, wyszukiwanie wykorzystujące synonimy lub tematyczną bazę wiedzy. Każde z tworzonych wewnątrz tego operatora zapytań może być złożone z kilku fragmentów za pomocą operatorów logicznych. Pod względem możliwości operator *CONTAINS* jest wręcz odpowiednikiem metody *Contains* w standardzie *SQL/MM*. Różnica pomiędzy nimi tkwi w tym, że operator *CONTAINS* w Oracle Text zwraca współczynnik zgodności dokumentu z zapytaniem i dodatkowo posiada zaawansowane mechanizmy pozwalające na sterowanie miarą zgodności. Funkcja *Score* również występuje w Oracle Text. Jej zadaniem jest jednakże tylko zwracanie współczynnika zgodności uzyskanego przez operator *CONTAINS*.

Możliwości Oracle Text nie kończą się na wspomnianej wyżej funkcjonalności. Istotną sferą podlegającą dalszemu rozwojowi są operacje klasyfikacji i klastrowania zbioru dokumentów. W celu umożliwienia wyżej wymienionej funkcjonalności Oracle Text oferuje specjalizowany typ indeksu *CTXRULE*. Kolejną możliwością znacznie wykraczającą poza standard *SQL/MM* jest wykorzystywanie podczas indeksacji i wyszukiwania struktury występującej w dokumentach tekstowych (HTML, XML). Pozwala to np. na wyszukiwanie słów lub fraz występujących w

określonym miejscu dokumentu np. w nagłówku (HTML) lub jako zawartość określonego elementu (XML).

Podsumowując, należy stwierdzić, że Oracle Text nie jest w pełni zgodny ze standardem SQL/MM, gdyż nie oferuje specjalizowanego typu danych do reprezentowania dokumentów, posiadającego metody zgodne ze standardem SQL/MM. Funkcjonalność i możliwości Oracle Text jednak znacząco wykraczają poza propozycje standardu i dotyczą głównie czterech dziedzin:

- miejsca składowania dokumentu,
- formatu dokumentu,
- możliwości klastrowania (grupowania) i klasyfikacji dokumentów,
- przetwarzania dokumentów semistrukturalnych takich jak XML czy HTML w sposób wykorzystujący ich wewnętrzną strukturę.

W przyszłości można byłoby sobie wyobrazić zbliżenie Oracle Text do standardu SQL/MM np. przez zastosowanie w Oracle Text specjalizowanego typu *FULLTEXT*. Mógłby pełnić on rolę wirtualnego dokumentu, którego zawartość byłaby uzyskiwana na podstawie wartości jego atrybutów. Atrybuty te mogłyby między innymi nieść ze sobą informację analogiczną do obecnych parametrów indeksów, takich jak *Datastore* i *Filter*. Ewolucja taka nie tylko byłaby zbliżeniem do ogólnie przyjętego standardu, ale także wyeliminowałaby pewne niepotrzebne zawiłości Oracle Text takie jak np. konieczność istnienia dodatkowych, sztucznych kolumn symulujących zawartość dokumentu, na których to zakładany jest indeks faktycznie odwołujący się do rzeczywistej zawartości dokumentu znajdującej się np. w tabeli zagnieżdżonej.

6.3. Oracle Spatial – obsługa danych przestrzennych

Kolejną opcją systemu zarządzania bazą danych Oracle pokrywającą część standardu SQL/MM jest Oracle Spatial [11]. Jest to opcja niestety niedostępna dla użytkowników Oracle10g w wersji Standard, a dla użytkowników wersji Enterprise dostępna za dodatkową opłatą. Powodem tego jest fakt, iż mimo możliwego szerokiego spektrum zastosowań, opcja ta wykorzystywana jest tylko przez instytucje państwowe i duże korporacje lub przedsiębiorstwa, dla których dane przestrzenne są podstawową informacją, której przechowywanie w systemie bazodanowym jest konieczne ze względu np. na jej zmienność w czasie.

W przypadku Oracle Spatial, tak jak w standardzie SQL/MM, mamy do czynienia ze specjalizowanym typem *SDO_GEOMETRY* pozwalającym na przechowywanie geometrii. Różnica w stosunku do standardu polega na tym, że typ jest jeden i pozwala na przechowywanie geometrii o dowolnym charakterze: punktu, linii, wielokąta, zbioru różnorodnych elementów. Standard SQL/MM przewiduje natomiast specjalizowane typy dla określonych typów geometrii. Informacja o zawartości obiektu typu *SDO_GEOMETRY* jest przechowywana w jednym z jego podstawowych atrybutów *SDO_GTYPE*. Wartością tego atrybutu jest liczba czterocyfrowa. Dwie ostatnie cyfry określają typ geometrii, z jakim mamy do czynienia. Typy geometrii pokrywają typy proponowane przez SQL/MM. Są to: punkt, linia, wielokąt (w tym koło), kolekcja elementów, zbiór punktów, zbiór linii i zbiór wielokątów.

Oracle Spatial nie wprowadza specjalnego rozróżnienia na typy obiektów zero, jedno lub dwuwymiarowych, natomiast bardzo istotną cechą geometrii jest to ile wymiarów zostało użytych do opisanie jej poszczególnych składowych, czy w jakim układzie współrzędnych geometria się znajduje. Informację tę opisuje pierwsza cyfra w atrybucie *SDO_GTYPE*. SQL/MM nie przewiduje definiowania obiektów trójwymiarowych. Oracle Spatial umożliwia definiowanie obiektów w przestrzeni trójwymiarowej, ale posiada stosunkowo słabe wsparcie, jeśli chodzi o operatory działające na tego typu geometriach, dlatego w praktyce wykorzystuje się również struktury dwuwymiarowe.

Analogicznie do standardu SQL/MM bardzo duży nacisk został położony na możliwość wykorzystania różnorodnych układów odniesienia. Oracle Spatial pozwala na skorzystanie z ponad tysiąca różnorodnych układów odniesienia. W obiekcie *SDO_GEOMETRY* układ odniesienia jest opisywany przez atrybut *SDO_SRID*.

Metody typu *SDO_GEOMETRY* ograniczają się do ekstrakcji informacji atrybutu *SDO_GTYPE* i nie obejmują operowania na geometriach, czy weryfikowania wzajemnej zależności tak jak jest to w

standardzie SQL/MM. Funkcjonalność metod standardu SQL/MM odnaleźć można analogicznie do Oracle Text w operatorach i funkcjach.

Operatorów i funkcji działających na geometriach jest kilkadziesiąt. Do swego działania wymagają one istnienia specjalizowanego indeksu *MDSYS.SPATIAL_INDEX* na analizowanym atrybucie. Można je podzielić na następujące grupy:

- Funkcje do odczytu miar i właściwości dla obiektów geometrycznych: *SDO_AREA*, *SDO_LENGTH* itp.
- Funkcje i operatory do określania zależności przestrzennych pomiędzy geometriami: *SDO_ANYINTERACT*, *SDO_CONTAINS*, *SDO_INSIDE*, *SDO_RELATE*, *SDO_GEOM.WITHIN_DISTANCE* itp.
- Funkcje do tworzenia nowych obiektów geometrycznych: *SDO_BUFFER*, *SDO_CENTROID*, *SDO_MBR*, *SDO_UNION*, *SDO_INTERSECTION* itp.
- Funkcje weryfikujące poprawność geometrii: *VALIDATE_GEOMETRY*, *VALIDATE_LAYER*,
- Funkcje agregujące: *SDO_AGGR_CENTROID*, *SDO_AGGR_CONCAT_LINES*, *SDO_AGGR_UNION* itp.

SQL/MM wprowadza także metody pozwalające na operacje bardzo częste i istotne w systemach geograficznych, a mianowicie konwersje układów odniesienia. W Oracle Spatial służą do tego funkcje i procedury pakietu *SDO_CS* np. *TRANSFORM*, *TRANSFORM_LAYER* etc.

Istotną informacją dotyczącą indeksowania danych przestrzennych jest to, że wymaga ona zapisania informacji o geometriach w tzw. metadanych. Dzięki temu zarówno operacja indeksowania jak i potencjalne aplikacje wykorzystujące geometrie mogą uzyskać informacje o obszarze zajmowanym przez geometrie, a także o dokładności, jaka jest konieczna podczas indeksacji i operowaniu na geometriach.

Bardzo ważnym rozszerzeniem Oracle Spatial, praktycznie niezauważonym przez SQL/MM, jest możliwość przechowywania pod postacią geometrii obiektów LRS (Linear Referencing System). Umożliwiają one na przechowywanie w geometriach nie tylko własności przestrzennych, ale także dodatkowych wartości o charakterze wag lub miar. Są one podstawowym elementem wykorzystywanym przez aplikacje transportowe.

Porównując metody standardu SQL/MM a funkcje i operatory Oracle Spatial odnosi się wrażenie pełnego pokrycia standardu oraz znacznego jego rozwinięcia. Czy możliwa jest ewolucja opcji Oracle Spatial do bardziej zbliżonej standardowi SQL/MM? Wydaje się, że tak, choć ujęcie funkcjonalności Oracle Spatial w zbiór typów SQL/MM wymagałoby znacznego ograniczenia obecnych możliwości Oracle Spatial.

7. Podsumowanie

SQL/MM jest stosunkowo nowym standardem, którego zadaniem jest specyfikacja standardowych bibliotek typów obiektowych SQL do obsługi zaawansowanych rodzajów danych, których konstrukcje języka SQL bezpośrednio nie wspierają. Obecnie SQL/MM, w zakresie obsługi zaawansowanych rodzajów danych zawiera specyfikacje dla danych przestrzennych, tekstowych i obrazów. Aktualna wersja serwera bazy danych Oracle: Oracle10g, Release 1 (10.1) implementuje jedynie specyfikację SQL/MM *Still Image*, oferując jednocześnie alternatywne mechanizmy do przetwarzania obrazów, wykraczające poza standard.

W zakresie zgodności Oracle10g ze specyfikacjami dla danych tekstowych i przestrzennych można w chwili obecnej mówić jedynie o zgodności funkcjonalnej, gdyż nazwy i składnia wykorzystywanych typów danych i funkcji są inne w Oracle10g i w standardzie. Należy jednak stwierdzić, że system zarządzania bazą danych Oracle10g posiada bardzo duże możliwości w zakresie wsparcia dla danych tekstowych i przestrzennych, uwzględniające funkcjonalność przewidzianą w standardzie SQL/MM, oferując dodatkowo wiele użytecznych mechanizmów nieobjętych standardem.

Standard SQL/MM, mimo że niekompletny ze względu na brak specyfikacji dla danych audio i wideo, już w tej chwili może służyć jako punkt odniesienia dla oceny funkcjonalności systemów zarządzania bazami danych dostępnych na rynku. Fakt implementacji jednej ze specyfikacji standardu SQL/MM przez jednego z liderów na rynku serwerów baz danych, jakim niewątpliwie jest Oracle,

daje nadzieję na to, że standard SQL/MM stanie się podstawą dla tworzenia przenaszalnych aplikacji operujących na zaawansowanych rodzajach danych, choć z pewnością jeszcze jest za wcześnie by mówić o pełnym sukcesie standardu..

Niniejszy artykuł skupia się na częściach standardu SQL/MM poświęconych obsłudze zaawansowanych typów danych, wspominając jedynie istnienie części szóstej standardu, poświęconej eksploracji danych. Jest to część najmłodsza, a ponadto nie jest ona jedyną próbą standaryzacji w dziedzinie eksploracji danych. Na razie więc trudno oceniać jej ewentualny wpływ na istniejące systemy eksploracji danych. Należy jednak zwrócić uwagę, że Oracle10g posiada opcję poświęconą eksploracji danych (Data Mining Option) [8], która obejmuje metody eksploracji przewidziane w standardzie SQL/MM, ale posiada interfejs inny niż przyjęty w standardzie SQL/MM.

8. Bibliografia

- [1] ISO/IEC 13249-1:2002, Information Technology – Database Languages – SQL Multimedia and Application Packages – Part 1: Framework. ISO, 2002.
- [2] ISO/IEC 13249-2:2003, Information Technology – Database Languages – SQL Multimedia and Application Packages – Part 2: Full-Text. ISO, 2003.
- [3] ISO/IEC 13249-3:2003, Information Technology – Database Languages – SQL Multimedia and Application Packages – Part 3: Spatial, ISO, 2003.
- [4] ISO/IEC 13249-5:2003, Information Technology – Database Languages – SQL Multimedia and Application Packages – Part 5: Still Image. ISO, 2003.
- [5] ISO/IEC 13249-6:2002, Information Technology – Database Languages – SQL Multimedia and Application Packages – Part 6: Data mining. ISO, 2002.
- [6] Melton J., Eisenberg A.: SQL Multimedia and Application Packages (SQL/MM). SIGMOD Record 30(4), 2001.
- [7] Melton J., Eisenberg A.: SQL:1999, formerly known as SQL3. SIGMOD Record 28(1), 1999.
- [8] Oracle Data Mining Concepts 10g, Release 1 (10.1). Oracle, 2003.
- [9] Oracle interMedia User's Guide 10g, Release 1 (10.1). Oracle, 2003.
- [10] Oracle interMedia Reference 10g, Release 1 (10.1). Oracle, 2003.
- [11] Oracle Spatial User's Guide and Reference 10g, Release 1 (10.1). Oracle, 2003.
- [12] Oracle Text Application Developer's Guide 10g, Release 1 (10.1). Oracle, 2003.
- [13] Stolze K.: SQL/MM Spatial: The Standard to Manage Spatial Data in Relational Database Systems. BTW 2003, 2003.