Andrzej CHOJNACKI[1]

MODELOWANIE ROZMYTE W ANALIZIE JAKOŚCIOWEJ Z WYKORZYSTANIEM ŚRODOWISKA OLAP

Streszczenie

W referacie przedstawiono matematyczny deterministyczny model struktury danych w środowisku OLAP w postaci wielowymiarowej i wielopoziomowej kostki. Model został uogólniony na przypadek danych o charakterze rozmytym lub interpretowanych z wykorzystaniem teorii zbiorów rozmytych i logiki rozmytej. Wskazano na możliwości wykonywania analiz jakościowych baz i hurtowni danych w oparciu o przedstawione modele.

Abstract

The article presents mathematical deterministic model of data structure applied in OLAP environments as multidimensional and multilevel cube. This model is generalized to the fuzzy data or data interpreted with fuzzy sets theory and fuzzy logic. Possibilities of quality analysis of data bases and data warehouses using presented models are described.

1 WPROWADZENIE

Współczesne hurtownie danych wyposażone są w nowoczesne narzędzia analityczne OLAP (*On-Line Analitycal Processing*) [2], które umożliwiają wykonywanie analiz wielowymiarowych [1]. Jest to możliwość konstruowania i analitycznego przetwarzania modelu wielowymiarowego danych, zorientowanego na procesy biznesowe.

Ekonomia i zarządzanie posługują się głównie językiem naturalnym, określeniami nieostrymi. Język ten cechuje też kadrę zarządzającą i decydentów w firmach. Modelowanie nieostrości w OLAP pozwala uzyskać rozmytą reprezentację danych, opartą na języku naturalnym, zrozumiałym dla wszystkich. Można to uzyskać dzięki zastosowaniu teorii zbiorów rozmytych.

[1] Dr hab. inż. Andrzej Chojnacki jest profesorem Warszawskiej Wyższej Szkoły Informatyki i Wojskowej Akademii Technicznej.

Analiza jakościowa oparta na modelowaniu rozmytym umożliwia dokonanie ocen i podsumowań w języku naturalnym, w formie zrozumiałej dla przeciętnego użytkownika [3]. Wykonana w środowisku OLAP stanowi proste i tanie rozwiązanie, umożliwiające podsumowanie dużych, trudno interpretowalnych liczbowych zbiorów danych w języku zgodnym z percepcją człowieka.

2 DETERMINISTYCZNY MODEL OLAP

Za pomocą technologii OLAP dane z hurtowni przekształca się do postaci wielowymiarowej, odmiennej od tradycyjnej, stosowanej w bazach danych, znormalizowanej struktury relacyjnej, ponieważ modele danych stosowane do projektowania systemów OLTP (On-Line Transaction Processing) nie nadają się do modelowania złożonych zapytań.

Ta nowoczesna technologia umożliwia modelowanie pozyskanych, wyczyszczonych i ujednoliconych danych do postaci zagregowanych, wielopoziomowych i wielowymiarowych struktur zwanych potocznie „kostkami", odzwierciedlających wielowymiarowy model działalności organizacji [3]. Następnie serwer OLAP udostępnia wielowymiarowe dane aplikacjom użytkowników. Interpretuje i przetwarza zapytania klientów. Struktury wielowymiarowe mogą być posadowione w dedykowanych, trwałych bazach wielowymiarowych, w tymczasowych kostkach rezydujących w pamięci lub w bazach relacyjnych.

Niech $\overline{N}$ oznacza liczbę wymiarów występujących w systemie OLAP. Wymiar n-ty ($n = \overline{1, N}$) opisany jest parą $\langle k_n, D_n \rangle$, gdzie k_n jest nazwą wymiaru, a D_n zbiorem jego możliwych wartości. Nazwa $k_n = \{d_n^e\}_{e=0}^{E_n}$ jest zbiorem nazw d_n^e semantycznie powiązanych, przy czym nazwa d_n^0 jest ogólną nazwą wymiaru (np. *lokalizacja*), natomiast pozostałe nazwy są jej konkretyzacją (np. *kraj*, *region*, *województwo*, *miasto*). Elementami zbioru D_n są więc ciągi $w_n = \langle w_n^0, w_n^1, w_n^2, ..., w_n^{E_n} \rangle$, w których $w_n^e \in D_n^e$, gdzie D_n^e jest zbiorem możliwych wartości nazwy d_n^e, czyli $D_n = \prod_{e=1}^{E_n} D_n^e$. Przyjmuje się, że $D_n^0 = \{ALL\}$, czyli zawsze $w_n^0 = ALL$. Dodatkowo z każdą nazwą d_n^e może być powiązany zbiór $Q_n^e \in Q_n$ atrybutów tej nazwy (np. dla nazwy *miasto* wymiaru *lokalizacja* zbiór ten może zawierać następujące atrybuty: *liczba mieszkańców*, *powierzchnia*). Każdy atrybut o nazwie q ze zbioru Q_n nazw wszystkich atrybutów n-tego wymiaru uwzględnianych w systemie OLAP może przyjmować wartości ze zbioru V_q.

W zbiorze k_n można zdefiniować relację binarną w ten sposób, że dwie różne nazwy z tego zbioru są ze sobą w tej relacji wtedy i tylko wtedy, gdy druga z nich jest semantycznie węższa niż pierwsza, czyli pierwsza jest hiperonimem w stosunku do drugiej, i odwrotnie druga jest hiponimem w stosunku do pierwszej. Oczywiście

nazwa d_n^0 jest hiperonimem w stosunku do każdej z pozostałych nazw, a relacja jest przeciwzwrotna, przeciwsymetryczna i przechodnia (tzw. relacja ostrego porządku). Tę relację można przedstawić w postaci acyklicznego, spójnego grafu skierowanego $H_n = \langle k_n, T_n \rangle$, gdzie $T_n \subseteq k_n \times k_n$ jest zbiorem takich łuków $\langle d_n^{e\prime}, d_n^{e\prime\prime} \rangle$, że nazwa $d_n^{e\prime\prime}$ jest hiponimem w stosunku do nazwy $d_n^{e\prime}$. Każda droga w tym grafie rozpoczynająca się w wierzchołku d_n^0 nazywana jest hierarchią n-tego wymiaru. Liczbę możliwych hierarchii oznaczmy symbolem I_n. Jest ona nie większa od 2^{E_n}. Niech $h_{ni} = \langle h_{ni}^0, h_{ni}^1, h_{ni}^2, ..., h_{ni}^{J_{ni}} \rangle$ oznacza hierarchię nr i n-tego wymiaru. Liczba J_{ni} nazywana jest liczbą poziomów i – tej hierarchii n - tego wymiaru, a największą z tych liczb – liczbą poziomów hierarchii n - tego wymiaru.

Hierarchia $h_{ni} = \langle h_{ni}^0, h_{ni}^1, h_{ni}^2, ..., h_{ni}^{J_{ni}} \rangle = \langle d_n^0, d_n^{e_n^{i_1}}, d_n^{e_n^{i_2}}, ..., d_n^{e_n^{i_{J_{ni}}}} \rangle$ w zbiorze D_n generuje ciąg o długości $J_{ni} + 1$ podziałów tego zbioru na rodziny $L_{ni}^j = \{L_{ni}^{ju}\}_{u \in U_{ni}^j}$ podzbiorów zbioru D_n, gdzie U_{ni}^j jest zbiorem indeksów j-tego podziału zbioru D_n $(j = \overline{0,J_{ni}})$. Podzbiory $L_{ni}^{ju} \subseteq D_n$ zdefiniowane są następująco:

– $L_{ni}^{0,1} = D_n$, przy czym $U_{ni}^0 = \{1\}$;

– jeśli $w'_n = \langle w'^0_n, w'^1_n, w'^2_n, ..., w'^{E_n}_n \rangle$ oraz $w''_n = \langle w''^0_n, w''^1_n, w''^2_n, ..., w''^{E_n}_n \rangle$, to w'_n i w''_n należą do tego samego podzbioru L_{ni}^{ju} wtedy i tylko wtedy, gdy jednocześnie należą do pewnego podzbioru $L_{ni}^{j-1u'}$ oraz $w'^{e_j^i}_n = w''^{e_j^i}_n$.

Widać, że hierarchia h_{ni} generuje dendryt $G_{ni} = \langle L_{ni}, W_{ni} \rangle$, którego wierzchołkami są wszystkie zbiory L_{ni}^{ju}, czyli $L_{ni} = \bigcup_{j=0}^{J_{ni}} \left\{ L_{ni}^{ju} \right\}_{u \in U_{ni}^j}$, a $\langle L_{ni}^{ju}, L_{ni}^{j+1u'} \rangle \in W_{ni}$ wtedy i tylko wtedy, gdy $L_{ni}^{j+1u'} \subseteq L_{ni}^{ju}$. Grafowi H_n obrazującemu wszystkie możliwe hierarchie n-tego wymiaru odpowiada więc graf $G_n = \langle L_n, W_n \rangle$ będący sumą dendrytów G_{ni} dla $i = \overline{1,I_n}$. Zbiorem L_n wierzchołków grafu G_n są wszystkie podzbiory L_{ni}^{ju}, czyli $L_n = \bigcup_{i=0}^{I_n} L_{ni}$, a łuki łączą te zbiory, z których następnik jest podzbiorem właściwym poprzednika i oba zbiory powstały z tej samej hierarchii.

Grafy H_n $(n = \overline{1,N})$ umożliwiają przeprowadzenie agregacji wymiarów. Niech $H = \langle k, T \rangle$ będzie grafem skierowanym, którego zbiorem k wierzchołków jest produkt kartezjański zbiorów k_n, tzn. $k = \prod_{n=1}^{N} k_n$, a para $\langle k'_n, k''_n \rangle \in T$ wtedy i tylko wtedy, gdy $k'_n \neq k''_n$ oraz dla każdego $n = \overline{1,N}$ para $\langle d'_n, d''_n \rangle$ jest łukiem w grafie H_n lub $d'_n = d''_n$.

Niech $D = \prod_{n=1}^{N} D_n$ będzie zbiorem wszystkich możliwych krotek wartości wymiarów. Podobnie jak powyżej w zbiorze D można przeprowadzić agregację grafów G_n w taki sposób, że otrzymuje się graf zagregowany $G = \langle L, W \rangle$, którego zbiór wierzchołków $L = \prod_{n=1}^{N} L_n$, natomiast łuk łączy dwa ciągi podzbiorów zbioru D wtedy i tylko wtedy, gdy są to ciągi różne oraz każdy podzbiór ciągu drugiego jest podzbiorem (niekoniecznie właściwym) odpowiadającego mu podzbioru ciągu pierwszego, czyli para takich podzbiorów jest łukiem w odpowiadającym im grafie G_n.

Charakterystyczną cechą systemu OLAP są miary, które pozwalają na wyznaczenie wskazanych wartości na podstawie znajomości elementów zbioru D. Niech K oznacza liczbę tych miar. Wartość każdej miary może być różna dla różnych wartości atrybutów rozpatrywanych w systemie OLAP. Niech $Q = \bigcup_{n=1}^{N} Q_n$ będzie zbiorem nazw wszystkich atrybutów, natomiast $V = \prod_{q \in Q} V_q$ – zbiorem wszystkich możliwych ciągów wartości atrybutów. Niech ponadto M_k oznacza zbiór wartości k-tej miary ($k = \overline{1,K}$). Każda miara opisana jest trójką $\langle m_k, M_k, F_k \rangle$, gdzie m_k jest nazwą k-tej miary (np. *sumaryczna wielkość sprzedaży*), natomiast F_k jest funkcją częściową, której wartością jest element zbioru M_k, tzn. $F_k : D \times V \to M_k$. Funkcja ta określona jest dla takich ciągów:

$$< \left\langle w_1^0, w_1^1, w_1^2, \ldots, w_1^{E_1} \right\rangle, \left\langle w_2^0, w_2^1, w_2^2, \ldots, w_2^{E_2} \right\rangle, \ldots, \left\langle w_N^0, w_N^1, w_N^2, \ldots, w_N^{E_N} \right\rangle, v^1, v^2, \ldots, v^Q >,$$

dla których v^q jest wartością co najmniej jednego atrybutu $q \in Q_n^e$ pewnej nazwy d_n^e przyjmującej wartość w_n^e. Niech $F = \langle F_1, F_2, \ldots, F_K \rangle$ będzie funkcją wektorową opisującą wszystkie miary. Wtedy modelem systemu OLAP jest tzw. kostka OLAP zdefiniowana następująco:

OLAP $= \langle D, H, \{M_k\}_{k=1}^{K}, F \rangle$.

Oczywiście konkretna realizacja kostki OLAP zawiera tylko takie krotki zbiorów D oraz V, które są elementami znanego zbioru danych. Można więc powiedzieć, że dla znanych zbiorów danych D oraz V fizyczna kostka OLAP jest kostką, w której zbiór D oraz funkcja F zostały „obcięte" do tych zbiorów. Ponadto przyjmuje się często, że wartości atrybutów zmiennych opisujących wymiary nie są brane pod uwagę przy konstruowaniu miar. Wtedy należy w powyższych rozważaniach przyjąć, że funkcja F jest funkcją stałą za względu na elementy zbioru V. W praktyce zakłada się także, że nie muszą występować wszystkie elementy kostki OLAP np. hierarchie. Wtedy przedstawiony model upraszcza się i zawiera tylko fragmenty obligatoryjne.

3 ROZMYTY MODEL OLAP

Współczesne bazy danych, w tym też hurtownie danych wyposażone w OLAP, przechowują informację precyzyjną. Bardziej efektywna i silniejsza reprezentacja wiedzy możliwa jest przy użyciu teorii zbiorów rozmytych. Inteligentne procedury przetwarzania informacji w OLAP mogą wtedy polegać na konwertowaniu (retranslacji) danych numerycznych do postaci lingwistycznej, a następnie na generowaniu analiz lingwistycznych, znajdowaniu wcześniej nieznanych zależności i schematów oraz wspieraniu procesu podejmowania decyzji w otoczeniu rozmytym [5].

Występujące w kostce ***OLAP*** nazwy w zbiorach $\bigcup_{n=1}^{N} k_n$, Q oraz $\{m_k\}_{k=1}^{K}$ mogą być interpretowane jako symbole zmiennych przyjmujących wartości odpowiednio ze zbiorów $\bigcup_{n=1}^{N}\bigcup_{e=0}^{E_n} D_n^e$, V oraz $\{M_k\}_{k=1}^{K}$. Mogą to być zmienne przyjmujące wartości deterministyczne np. liczbowe lub nazwy własne, ale mogą to też być zmienne lingwistyczne [7] tzn. takie zmienne, których wartościami są określenia w języku naturalnym zwane terminami lingwistycznymi (np. dla zmiennej lingwistycznej *sprzedaż* mogą to być wartości *mała*, *duża, bardzo duża*). Przyjmijmy, że R spośród wymienionych wyżej nazw są to zmienne lingwistyczne stanowiące zbiór $Z = \{Z_r\}_{r=1}^{R}$, a zmiennej lingwistycznej Z_r odpowiada skończony zbiór terminów lingwistycznych $E_r = \{t_r^s\}_{s=1}^{S_r}$ oraz zbiór możliwych jej wartości fizycznych, który oznaczmy symbolem Z_r (np. dla zmiennej *sprzedaż* może to być *zbiór liczb rzeczywistych nieujemnych*). Oznacza to, że jeśli w kostce ***OLAP*** zmienna lingwistyczna Z_r jest nazwą d_n^e elementu wymiaru, to $Z_r = D_n^e$, jeśli jest to nazwa atrybutu q, to $Z_r = V_q$, a jeśli jest to nazwa m_k miary, to $Z_r = M_k$. Jako znaczenie terminu lingwistycznego t_r^s przyjmuje się pewien zbiór rozmyty E_r^s. Wtedy $E_r^s = \{\langle t, \mu_r^s(t)\rangle : t \in Z_r\}$, gdzie $\mu_r^s : Z_r \to [0;1]$ jest funkcją przynależności zbioru rozmytego E_r^s. Uwzględnienie rozmytości w modelu OLAP wymaga więc rozszerzenia kostki ***OLAP*** o zbiory E_r dla wszystkich nazw, których rozmytości w modelu będą uwzględniane, to znaczy dla wszystkich nazw będących zmiennymi lingwistycznymi ze zbioru Z. Ponadto kostka musi być rozbudowana o zbiory rozmyte E_r^s dla wszystkich terminów lingwistycznych t_r^s tych zmiennych lingwistycznych. Rozmyta kostka OLAP może być więc zdefiniowana jako

$$\boldsymbol{OLAP}_{\boldsymbol{rozm}} = \left\langle D, H, \{M_k\}_{k=1}^{K}, F, Z, \{E_r\}_{r=1}^{R}, \left\{\left\{E_r^s\right\}_{s=1}^{S_r}\right\}_{r=1}^{R}\right\rangle .$$

Oczywiście rozpatrywane w kostce $\boldsymbol{OLAP}_{rozm}$ zbiory rozmyte E_r^s są ze sobą powiązane zależnościami zdefiniowanymi w kostce. Powiązania takie wynikają w szczególności z:

- grafów $\{H_n\}_{n=1}^N$ powiązań nazw wymiarów,
- zbiorów $\{Q_n\}_{n=1}^N$ nazw atrybutów powiązanych z nazwami wymiarów,
- funkcji F opisującej miary i określonej na zbiorach odpowiadających wymiarom i atrybutom.

W rezultacie tylko niektóre funkcje μ_r^s przynależności zbiorów rozmytych muszą być definiowane przy konstruowaniu kostki ***OLAP***. Pozostałe są odpowiednimi złożeniami tych zbiorów lub też definiowane mogą być innymi metodami pośrednimi np. poprzez zastosowania kwantyfikatorów lingwistycznych. Otrzymujemy w ten sposób możliwość interpretacji wartości deterministycznych w kategoriach nieprecyzyjnego języka naturalnego.

4 PODSUMOWANIE

Rozmyty model OLAP, oparty na zmiennych lingwistycznych, umożliwia wykonanie analiz jakościowych w zakresie:

1. Rozszerzenia klasycznego formułowania warunków zapytań poprzez stosowanie języka naturalnego. Użytkownik ma możliwość konstruowania tzw. zapytań nieprecyzyjnych [8], czyli zapytań do bazy danych OLAP z użyciem terminów lingwistycznych. Uzyska w ten sposób wiedzę na temat stopnia dopasowania danych do zapytania za pomocą stopni przynależności danych do zbioru rozmytego, w szczególności stopień spełnienia warunków zapytania nieprecyzyjnego dla różnych terminów lingwistycznych.
2. Lingwistycznego podsumowania danych [8] – opisu zawartości danych w kostce OLAP za pomocą zwięzłych zdań wyrażonych w języku naturalnym na różnych poziomach agregacji danych. Np. „W styczniu wystąpiła wysoka sprzedaż lodówek", „W 2006 r. odnotowaliśmy niską sprzedaż dla klientów z województwa lubelskiego". Dla tak wyrażonych sformułowań oblicza się stopień prawdy czyli wartość funkcji przynależności.
3. Eksploracji danych – odkrywania lingwistycznych reguł asocjacyjnych opisujących związki pomiędzy wybranymi wartościami poszczególnych zmiennych lingwistycznych [5].

Literatura

1. Agrawal R., Gupta A., Sarawagi S.: *Modeling Multidemnsional Databases.* In Int. Conf. On Dta Engineering. IEEE 1997.
2. Blaschka M., Sapia C., Höffing G.: *Finding Your Way through Multidemnsional Data Models.* [w:] Proceedings Int. Workshop on Data Warehouse Design and OLAP Technology, Vienna, August 1998.
3. Chaudhuari S., Dayal U.: *An Overview of Data Warehousing and OLAP Technologies. ACM SIGMOND* Rekord 26 (1), Marc 1997.
4. Chojnacki A., Borzęcka H.: *Assesment of Economic Activity of the Company on the Base of Fuzzy Inference Rules*. [w:] Studia Informatica 1 (12)2009, PL ISSN: 1731-2264.
5. Chojnacki A., Borzęcka H.: *Fuzzy Modeling for OLAP Quality Analysis*. [w:] Proceedings of Artificial Intelligence Studies, Vol.6 (29)/2009, Publishing House of University of Podlasie, ISBN 978-83-7051-525-6, Proceedings on International Conference on Artificial Intelligence AI-24'2009.
6. Kacprzyk J., Zadrożny S.: *Data Mining via Linguistic Summaries of Databasees: anIinteractive Approach.* Ding L. (red.): A New Paradigm of Knowledge Engineering by Soft Computing, ss 325-345. Singapore: World Scientific, 2001.
7. Zadeh L. A.: *The Concept of a Linguistic Variable and Its Application to Approximate Reasoning*. Part I-III. Information Sciences, 8,9:199-249,301-357,43-80, 1975.
8. Zadrożny S.: *Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych*. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2006.