GENERALIZED NET MODEL OF THE PROCESS OF DATAWAREHOUSE

Daniela Orozova

Abstract: *A theoretical model, describing interconnection between datawarehouse components, is proposed. The sequence of actions, referring to creation and functioning of datawarehouse, as well the information flow, is defined. The model is universal and independent of datawarehouse's problem area.*

Key words: *Knowledge Discovery in Databases, Generalized nets, Datawarehouse, OLAP tools,, Intelligent database system.*

МОДЕЛ НА ПРОЦЕСИТЕ В СКЛАД ОТ ДАННИ ЧРЕЗ СРЕДСТВАТА НА ОБОБЩЕНИТЕ МРЕЖИ*

Даниела Орозова

Абстракт: *Създаден е теоретичен модел на процесите на взаимодействие на отделните компоненти на склад от данни (Datawarehouse). Определени са последователностите от действия, свързани със създаване и функциониране на склад от данни и информационните потоци в него. Даденият подход е универсален и не зависи от конкретната предметната област на склада.*

Ключови думи: *Откриване на знания в бази от данни, Обобщени мрежи, Склад от данни, OLAP средства, Интелигентни бази от данни.*

1. Архитектура на склад от данни

Огромното количество събрани данни значително превишава възможността на човека те да бъдат ефективно използвани без помощта на специализирани мощни средства за анализ на данни. Концепцията за склад от данни се изразява в преобразуването на големи масиви от данни в значима информация, която подпомага взимането на бизнес решения. Дефиниция за понятието „склад от данни” е дадена от Уилям Инмон през 1990 г., според която: „складът от данни е предметно-ориентиран, интегриран, зависещ от времето и неизменящ се набор от данни в подкрепа на процеса на взимане на решения”.

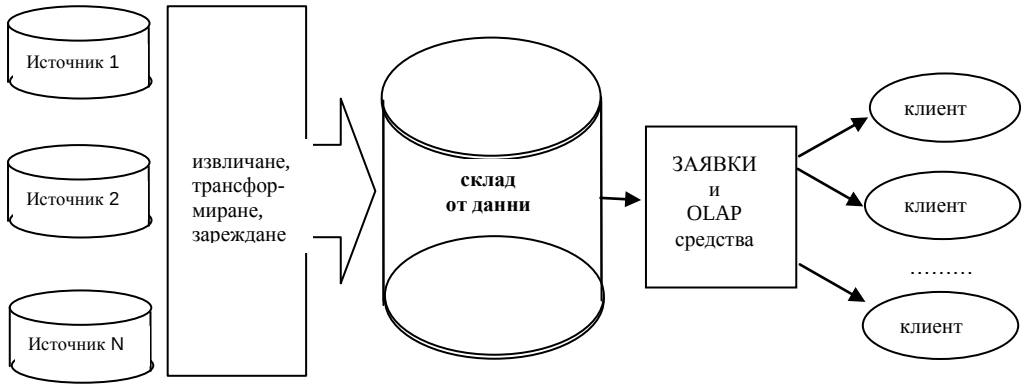
Повечето бази от данни се създават, за да се управлява реалния процес на организациите. Тези данни са **операционални**, а системите, които използват данните са **OLTP** (On-Line Transaction Processing). Системите за обработка на транзакции - складират големи количества данни и дават възможност на потребителите да достигат до данните чрез бърз, интерактивен (онлайн) достъп. От друга страна един склад с данни обикновено не съхранява текуща информация за дадена бизнес дейност, а данни удобни за колективна обработка. Той обединява данни от множество транзакционни системи, организира ги и прави данните достъпни за целите на анализа. Този тип данни

* Изследването е свързано с работата по проект Д12/21.02.2013, финансиран от фонд „Научни изследвания” при Бургаски свободен университет.

могат да се нарекат **информационни**, а системите за обработка на информационни данни - **OLAP** (On-Line Analytical Processing).

Процесът „Knowledge Discovery in Databases” (откриване на знания) включва подготовка на данните, избор на информативни признаци, пречистване на данните, приложение на метод за извличане и откриване на данните, преработка на данните и интерпретация на получените резултати [4]. Основен акцент в процеса е извличане и откриване на данните (Data Mining). **Data mining** също се използва за анализ на данните, но обхваща технологии, позволяващи да се открият неявни шаблони (образци) и зависимости в данните. Той позволява получаването на знания, като правила, описващи връзката между свойствата на данните, модели на данните, резултатите от класификацията и клъстеризацията на данните.

Схематично представяне на процеса на конструиране на склад от данни е дадено на Фигура 1. Един **склад с данни** е хранилище на информация, събрана от множество източници, съхранена под обща (единна) схема на данните и която обикновено се съхранява на едно място. Складовете с данни се конструират чрез процеси на извличане, пречистване на данните, трансформация на данните, интегриране на данните, зареждане на данните и периодично обновяване на данните. Предоставят се средства за извличане на различна информация като *OLAP* средства, софтуер за *data mining* и различни системи, подпомагащи вземането на решения



Фигура 1. Схема на процеса на конструиране на склад от данни

2. Модел на процесите в склад от данни

Обобщено-мрежовият модел [1,2,3] на процесите на взаимодействие на отделните компоненти на склад от данни е представен на Фигура 2. Моделът е изграден от 6 прехода и 20 позиции, като преходите представят описаните по-долу процеси.

Първоначално и по време на цялото функциониране на обобщено мрежовият модел в позиция l_{17} присъства α-ядро с начална и текуща характеристика:

”*склад от данни*”.

В даден момент от време α-ядрото може да се разцепи на две ядра, като оригиналното ядро остава в първоначалната си позиция през цялото време. Всички постъпващи в позиция l_{17} ядра се сливат с оригиналното α-ядро.

Входни позиции за мрежата са: l_1 , l_2 *и* l_3 .

През позиция l_1 , в модела посъпват α-ядра, с начална характеристика:

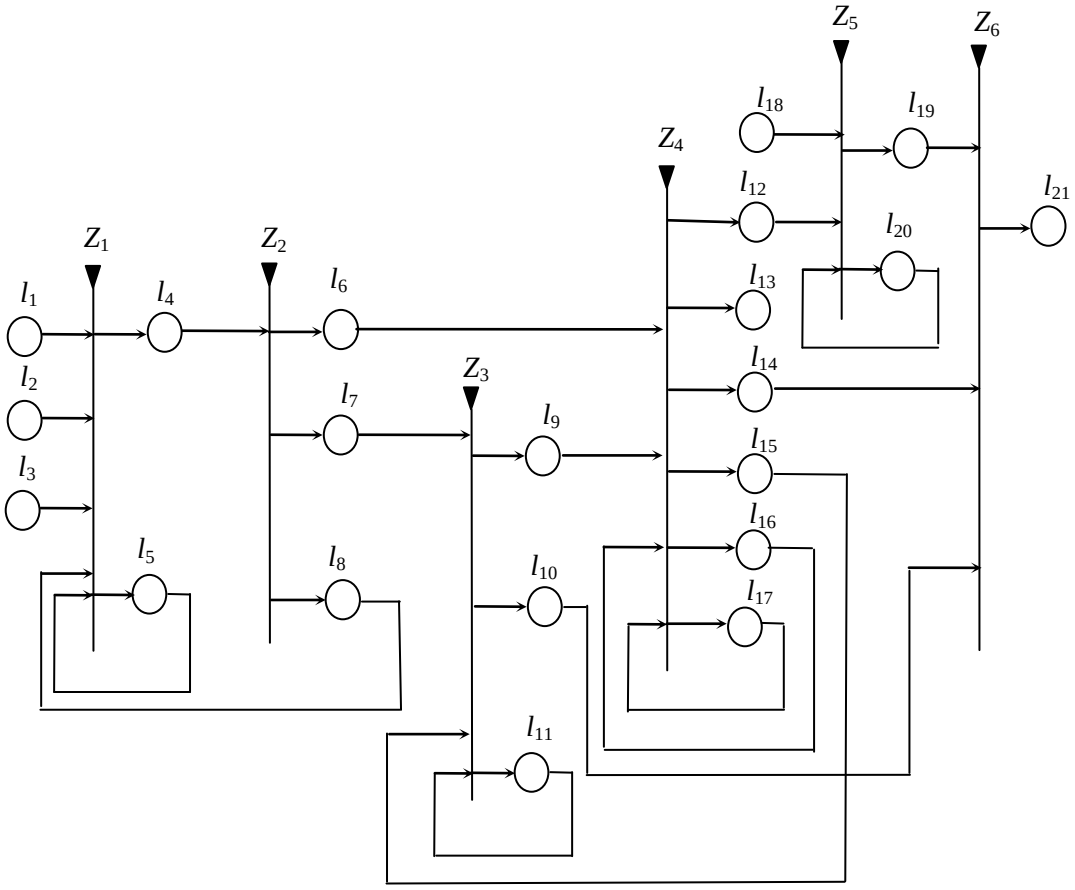
„*данни от релационни бази от данни”,*

през позиция l_2 посъпват α-ядра, с начална характеристика:

„*нерелационни данни от външни системи”,*

през позиция l_3 , посъпват α-ядра, с начална характеристика:

”*външни маркетингови данни*”



Фигура 2. Обобщено-мрежови модел на процесите на изграждане на склад от данни

Преходът Z_1 представя в модела извличането на данни от различни външни източници. От тези източници обикновено постъпват херерогенни данни от бази от данни, файлови системи, web-данни, e-mail и др. Поради това тук са необходими добре развити инструменти за достъп до различни източници от данни.

$$Z_1 = <\{ l_1 , l_2 , l_3 , l_5 , l_8 \}, \{ l_4, l_5 \}, r_1, \vee_1>,$$

$$r_1 = \begin{array}{c|cc} & l_4 & l_5 \\ \hline l_1 & F & T \\ l_2 & F & T \\ l_3 & F & T \\ l_5 & W_{5,4} & T \\ l_8 & F & T \end{array}$$

където предикатът $W_{5,4}$= “Извлечени са данни от външен източник”.

Условието за настъпване на прехода е: $\vee_1 = \vee\ (l_1 , l_2 , l_3 , l_5 , l_8)$.
След активиране на прехода в позиция l_4 постъпва α-ядро, с характеристика:

„*операционални данни”,*

в позиция l_5 , стоиа α-ядра, с характеристика:

„*архив* с *операционални данни”,*

Преходът Z_2 представя преработването на данните, които ще се съхраняват в хранилището: пречистване, филтриране и обобщаване на данните на различни нива на обобщение. Данните могат да съдържат пропуски, шум, аномални стойности, неподходящи или неизползваеми данни, да са недостатъчни и т.н. извършва се проверка и отстраняване на противоречиви или дублирани данни. От операционалните данни трябва да се създадат обощени данни – групирани по различни категории и за всяка категория се съхранят важните характеристии. Понякога се налага данните да се преформатират и декомпозират като се понижава или повишава размерността на изходното пространство, прилагайки специални алгоритми.

$$Z_2 = <\{l_4\}, \{l_6, l_7, l_8\}, r_2, \vee_2>,$$

$$r_2 = \begin{array}{c|ccc} & l_6 & l_7 & l_8 \\ \hline l_4 & W_{4,6} & W_{4,7} & W_{4,8} \end{array}$$

където:
$W_{4,6}$ = “Данните са преработени, готови за постъпване в склада от данни”,
$W_{4,7}$ = “Необходимост от обобщаване и OLAP анализ на данните ”,
$W_{4,8}$ = “Заявка за допълнително извличане на данни от външни източници”,

Условието за настъпване на прехода е: $\vee_2 = \vee\ (l_4)$.
α-ядрата, постъпващи в позиции l_6, l_7 и l_8 получават следните характеристики:

- в позиция l_6: “*данни за въвеждане в склада от данни”;*
- в позиция l_7 “*заявка за анализ на данните към OLAP сървъра”;*
- в позиция l_8: “*заявка за допълнително извличане на данни*”.

Преходът Z_3 представя процеса на работа на OLAP сървъра. Данните се преобразуват в подходящ формат и се подлагат на анализ. OLAP сървъра извършва сложен статистически анализ на данните в реално време. Средствата позволяват на квалифицирани потребители да анализират данните по няколко параметъра, с цел изготвяне на пазарни прогнози, планиране на технически мощности и др. Те работят с “многомерни БД”, което предполага поддържане в склада от данни на групирани структури във вид на кубове от данни. Създаването и различните операции с тези структури от данни се реализират от OLAP сървъра.

$$Z_3 = <\{l_7, l_{11}, l_{15}\}, \{l_9, l_{10}, l_{11}\}, r_3, \vee_3>,$$

$$r_3 = \begin{array}{c|ccc} & l_9 & l_{10} & l_{11} \\ \hline l_7 & F & F & T \\ l_{11} & W_{11,9} & W_{11,10} & T \\ l_{15} & F & F & T \end{array}$$

където
$W_{11,9}$ = “Изпълнена е заявка за OLAP анализ на данните, преизчислените данни са готови за запис в склада от данни”.

$W_{11,9}$ = “Изпълнена е заявка за OLAP анализ на данни на краен потребител”.

Условието за настъпване на прехода е: $\vee_3 = \wedge(\vee(l_7, l_{15}), l_{11})$.
През позиция l_7 или l_{15} в модела постъпва α- ядро с характеристика:
“*заявка за анализ на данните към OLAP сървъра”.*
α-ядрата, постъпващи в позиции l_9 и l_{10} получават характеристиките съответно:
“ резултат от OLAP анализ на данните за въвеждане в склада от данни”;
“ резултат от OLAP анализ на данните за краен потребител”.
α-ядрото в позиции l_{11} не получава нова характеристика.

Преходът Z_4 представя входно-изходния информационен поток и съхранение на данните в склада от данни. Детайлните данни могат да се обобщават на различни нива и върху тях да се извършват различни OLAP операции; върху данните могат да се прилагат различни потребителски заявки за селекция, проекция, съединение, резюмиране на данните; данните могат да се конвертират в различни формати (електронни таблици, схеми, диаграми, отчети и др.); да се пакетират и остарелите данни да се съхраняват на външни носители.

$$Z_4 = <\{l_6, l_9, l_{17}\}, \{l_{12}, l_{13}, l_{14}, l_{15}, l_{16}, l_{17}\}, r_4, \vee_4>,$$

$r_4 =$

	l_{12}	l_{13}	l_{14}	l_{15}	l_{16}	l_{17}
l_6	F	F	F	F	$W_{6,16}$	$W_{6,17}$
l_9	F	F	F	F	$W_{9,16}$	$W_{9,17}$
l_{17}	$W_{17,12}$	$W_{17,13}$	$W_{17,14}$	$W_{17,15}$	T	T

където:
$W_{6,16}$ = “Постъпили са данни за запис в склада от данни”;
$W_{6,17} = W_{6,16}$;
$W_{9,16}$ = “Постъпил е резултат от OLAP заявка за запис в склада от данни”;
$W_{9,17} = W_{9,16}$;
$W_{17,12}$ = “Заявка за Data Mining анализ върху данни от склада”;
$W_{17,13}$ = “Извършен е запис на данни от склада на външен носител”;
$W_{17,14}$ = “Заявка за извличане на данни от краен котребител”;
$W_{17,15}$ = “Заявка към OLAP сървър за анализ на данни от склада”;

Условието за настъпване на прехода е: $\vee_4 = \wedge(\vee(l_6, l_9), l_{17})$.
След активиране на прехода:
α-ядрата, постъпващи в позиции l_{12} получават характеристика:
”данни от склада. заявка за Data Mining анализ”.
α-ядрата, в позиции l_{13} получават характеристика:
“ данните от склада, записани на външен носител”;
α-ядрата, в позиции l_{14} получават характеристика:
“данни от склада за предоставяне на краен потребител”;
α-ядрата, в позиции l_{15} получават характеристика:
“ данни от склада. заявка за OLAP анализ на данните”;
α-ядрата, постъпващи в позиции l_{16} се сливат с α-ядрото, което цикли в тази позизия по време на цялото функциониране на мрежата с характеристика:
"Хранилище на метаданни”;

α-ядрата, постъпващи в позиции l_{17} се сливат с α-ядрото, което цикли в тази позизия по време на цялото функциониране на мрежата с характеристика:

"”Склад от данни”.

Преходът Z_5 представя Data Mining анализи, прилагани върху данните от склада. Прилагат се различни видове анализ като: класификационен анализ, регресионен анализ, характеристичен анализ, клъстерен анализ, еволюционен анализ и др. За различните цели на анализа могат да се прилагат различни техники и средства за откриване на закономерности върху данните [4], например: невронни мрежи, дървета на решенията, алгоритми за клъстеризация, установяване на асоциации и т.н. Откритите закономерности се оценяват с помощта на различни коефициенти за измерване на приложимостта им и в резултат могат да се определят като знания.

През позиции l_{18} постъпва β-ядра с начална характеристика:

”*Data Mining средства*”.

Условието за настъпване на прехода е: $V_5 = \wedge\ (l_{12}, l_{18})$.
След активиране на прехода α-ядрото от позиции l_{12} се слива с β-ядро от позиции l_{18} и постъпва в позиция l_{20} за прилагане на избрано Data Mining средство (средство за извличане на знания от данните).

$$Z_5 = <\{l_{12}, l_{18}, l_{20}\}, \{l_{19}, l_{20}\}, r_5, V_5>,$$

$r_5 =$

	l_{19}	l_{20}
l_{12}	F	T
l_{18}	F	T
l_{20}	$W_{20,19}$	T

където:
$W_{20,19}$ = “Получени са нови знания, т.е. извлечени са модели (шаблони) на данните”.

α-ядрото, постъпващо в позиция l_{19} получава характеристика:

“ модели (шаблони) на данните”.

β-ядрото цикли в позиция l_{20} с текуща характеристика:

“ Data Mining (алгоритми) средства ”.

Интерпретацията на прехода Z_6 е представяне на получените резултати и знания на крайни потребители или бизнес приложенията. Тук се включва и визуализация на информацията с цел потребителят да разбере и интерпретира резултатите. Получената информация може да подпомага взимането на неструктурирани, стратегически решения.

$$Z_6 = <\{l_{10}, l_{13}, l_{14}\}, \{ l_{21}\}, r_6, V_6>,$$

$r_6 =$

	l_{21}
l_{10}	$W_{10,21}$
l_{13}	$W_{13,21}$
l_{14}	$W_{14,21}$

където
$W_{10,21}$ = “Данни, получени в резултат на OLAP анализ”;
$W_{13,21}$ = “Данни, получени в резултат на заявка към склада от данни”;

$W_{10,21}$ = “Знания, изведени в резултат на заявка към Data Mining средства”.

Условието за активиране на прехода е: $V_6 = \vee$ (l_{10}, l_{13}, l_{14}).

α_{13}-ядрото, постъпващо в позиция l_{21} получава характеристика

“*Резултат от потребителска заявка*”.

Прилагайки йерархичен оператор H_3 от теорията на обобщените мрежи [2], представения ОМ модел от фигура 2 може да бъде детайлизиран и разширен. Следващият модел, представен на фигура 3 може да се разглежда като подмрежа, заменяща позиция l_{20} в модела на функциониране на склад от данни.

Следва кратко представяне на ОМ модел на процеса на прилагане на Data Mining средства, подробно представен в [6]. Подмрежата съдържа 5 прехода и 21 позиции, групирани в две групи и свързани с два типа ядра, които ще постъпват в съответните типове позиции: α-ядра и l-позиции представят процеса на прилагане на Data Mining средства, β-ядра и t-позиции представят критериите за ограничаване на средствата и избор на подходящи Data Mining средства.

За краткост ще се използва означението α- и β-ядра вместо α_i- и β_j-ядра, където i, j са номерата на съответните ядра.

Първоначално едно β_0-ядро стои в позиция t_6 с начална характеристика

„*налични Data Mining средства*“.

На следващия преход от функционирането на мрежата β-ядрото се разделя на две. Оригиналното β-ядро ще продължи да стои в позиция t_6, докато другото β-ядро ще се придвижи към прехода Z_5, преминавайки през прехода Z_2.

Ядрата α_0 и α_1, постъпващи в мрежата през позиции l_0 и l_1, получават характеристики съответно: „*начални хипотези*“ и „*начални данни*“.

Ядрата β_1 и β_2 постъпват в мрежата през позиции t_0 и t_1 съответно. Тези ядра получават съответно начални характеристики:

„*нова Data Mining техника* (*средство*)“;

„*критерии за избор на Data Mining средства*“.

Следва кратко описание на отделните преходи.

$Z_1 = \langle\{l_0, l_1, l_{10}, t_2\}, \{l_2, l_3, l_4\}, r_1, \vee(\wedge(\, l_0, l_1), l_{10}, t_2)\rangle$;

$r_1 =$

	l_2	l_3	l_4
l_0	*False*	*False*	*True*
l_1	*False*	*False*	*True*
l_{10}	$W_{10,2}$	$W_{10,3}$	*False*
t_2	*False*	*True*	*False*

,

където:

$W_{10,2}$ = „Data Mining техниката е приложена“;

$W_{10,3} = \neg\, W_{10,2}$.

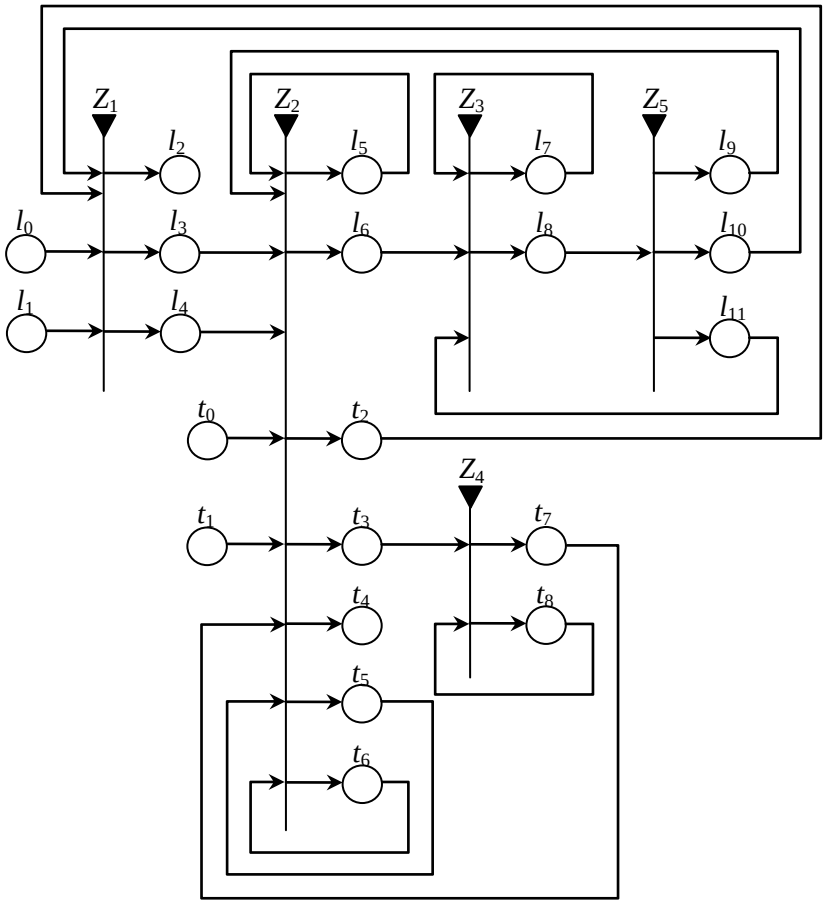
Като първо действие на прехода Z_1 α_0- и α_1-ядрата, които постъпват през позиция l_4 (от позиции l_0 и l_1), се сливат в едно ново α-ядро с характеристика

„*начални хипотези, начални данни*“.

Характеристиката на β-ядрото, което постъпва в позиция l_3 (от позиция t_2), получава характеристиката:

„*цел, Data Mining техника*“.

При следващото активиране на прехода Z_1 β-ядрата могат да постъпят в позиция l_2 или l_3. β-ядрото, което постъпва в позиция l_2, не получава нова характеристика.



Фигура 3. ОМ модел на процеса на прилагане на Data Mining средства

$Z_2 = \langle\{l_3, l_4, l_5, l_9, t_0, t_5, t_6, t_7, t_1\}, \{l_5, l_6, t_2, t_3, t_4, t_5, t_6\}, r_2, \vee(\, l_3, l_4, l_5, l_9, t_0, \wedge(\, t_1, t_5), t_6, t_7)\rangle$;

$r_5 =$

	l_5	l_6	t_2	t_3	t_4	t_5	t_6
l_3	*True*	*False*	*False*	*False*	*False*	*False*	*False*
l_4	*False*	*False*	*False*	*False*	*False*	*False*	*True*
l_5	$W_{5,5}$	$W_{5,6}$	*False*	*False*	*False*	*False*	*False*
l_9	*True*	*False*	*False*	*False*	*False*	*False*	*False*
t_0	*False*	*False*	*False*	*False*	*False*	*False*	*True*
t_5	*False*	*False*	*False*	*True*	*False*	*False*	*False*
t_6	*False*	*False*	*False*	*False*	$W_{6,4}$	$W_{6,5}$	*True*
t_7	*False*	*False*	*True*	*False*	*False*	*False*	*False*
t_1	*False*	*False*	*False*	*True*	*False*	*False*	*False*

където:

$W_{5,5}$ = „има Data Mining техника, която още не е приложена“;

$W_{5,6}$ = „Data Mining техниката е извлечена“;

$W_{6,4}$ = „Data Mining техниката е отхвърлена“;

$W_{6,5}$ = „Data Mining техниката е избрана“.

В позиция l_3 и l_4 α-ядрата получават характеристики съответно

„*избрани подходящи Data Mining техники, критерии за избор на Data Mining техники*“;

„*отхвърлена Data Mining техника*“.

β_1-ядрото, което постъпва в прехода Z_2 (от позиция t_0), ще се слее с оригиналното β_0-ядро, което стои в позиция t_6. β_3-ядрото, което постъпва в позиция t_5, получава характеристика

„*избрани Data Mining техники*“.

β-ядрата, които постъпват в позиции t_2 (от позиция t_7) и t_3 (от позиция t_5), не получават нова характеристика.

$$Z_3 = \langle\{l_6, l_7, l_{11}\}, \{l_7, l_8\}, r_3, \vee(l_6, l_7, l_{11})\rangle;$$

$$r_3 = \begin{array}{c|cc} & l_7 & l_8 \\ \hline l_6 & True & False \\ l_7 & W_{7,7} & W_{7,8} \\ l_{11} & True & False \end{array},$$

където:

$W_{7,7}$ = „има следващи стъпки от работата на текущата Data Mining техника“;

$W_{7,8}$ = „избрана е следващата стъпка от работата на текущата Data Mining техника“.

α-ядрото, което постъпва в позиция l_7, не получава нова характеристика, докато α-ядрото, което постъпва в позиция l_8, получава характеристиката

„*текуща стъпка от прилагането на избраната Data Mining техника*”

$$Z_4 = \langle\{t_3, t_8\}, \{t_7, t_8\}, r_4\rangle;$$

$$r_4 = \begin{array}{c|cc} & t_7 & t_8 \\ \hline t_3 & False & True \\ t_8 & W_{8,7} & W_{8,8} \end{array},$$

където:

$W_{8,7}$ = „избрана е Data Mining техника за поставената цел“;

$W_{8,8} = \neg W_{8,7}$.

β-ядрата, които постъпват в позиция t_8, не получават нова характеристика, докато β-ядрата, които постъпват в позиция t_7, получават характеристиката

„*цел, избрани Data Mining техники*“.

$$Z_5 = \langle\{l_8\}, \{l_9, l_{10}, l_{11}\}, r_5\rangle;$$

$$r_5 = \begin{array}{c|ccc} & l_9 & l_{10} & l_{11} \\ \hline l_8 & W_{8,9} & W_{8,10} & W_{8,11} \end{array},$$

където:

$W_{8,9}$ = „има следваща Data Mining техника, която може да бъде прилагана“;

$W_{8,10}$ = „последната възможна Data Mining техника е приложена“;

$W_{8,11}$ = „има следваща стъпка от текущата Data Mining техника, която ще се изпълнява”.

α-ядрата, които постъпват в позиции l_9 и l_{11}, не получават нови характеристики, докато α-ядрата, постъпващи в позиция l_{10}, получават характеристиката

„*цел, Data Mining средство, оценка от работата на Data Mining средството*“.

3. Приложения

Създаденият обобщено-мрежови модел показва връзките, начина на взаимодействие и процесите между отделните компоненти в склад от данни. Поетапно се проследяват информационните потоци в склада: вливане на данните, получавани от

различни хетерогенни източници, тяхното съхранение, описание и метаданни, архивиране, анализ и предоставяне на данни и знания на крайните потребителите или бизнес приложения. Моделът може да бъде детайлизиран чрез прилагане на йерархичен оператор H_3, който заменя преход или позиция от мрежата с нова мрежа, описваща по-детайлно съответен процес. Такъв модел намира приложение при създаване, изследване и оптимизиране на специализиран софтуер за работа със складове от данни.

Литература:

[1] Atanassov, K., Generalized Nets, World Scientific, Singapore, 1991.
[2] Atanassov, K. On Generalized Nets Theory. Prof. M. Drinov Academic Publ. House, Sofia, 2007.
[3] Atanassov, K., H. Aladjov. Generalized Nets in Artificial Intelligence. Vol.2: Generalized Nets and Machine Learning. “Prof. M. Drinov” Academic Publishing House, Sofia, 2001.
[4] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 1996.
[5] Shannon, A., E.Sotirova, D.Orozova, Geneneralized Net Model of Using Data Mining Techniques for Process of Undergraduate Matriculation in a Digital University, Eleventh Int. Workshop on GNs and Second Int. Workshop on GNs, IFSs, KELondon, 9-10 July 2010, 1-6.
[6] Sotirova, E., D. Orozova, Generalized Net Model of the Phases of the Data Mining Process, Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics, Volume II: Applications, IBS PAN – SRI PAS, Polish Academy of Sciences, Warsaw, 2010, pp.247-260.
[7] Sumathi, S., S.N. Sivanandam, Introduction to Data Mining Principles and its Applications, Studies in Computational Intelligence, Springer, Vol. 29, 2006.