

ЕКСПЕРИМЕНТАЛНО ИЗСЛЕДВАНЕ ЕФЕКТИВНОСТТА НА МОДЕЛНИЯ ФОРМАЛИЗЪМ СЪЩНОСТ - ВРЪЗКА – АТРИБУТИ

EXPERIMENTAL INVESTIGATION OF MODELING FORMALISM ENTITY – RELATIONSHIP – ATTRIBUTES EFFECTIVENESS

ЕКСПЕРИМЕНТАЛНОЕ ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МОДЕЛЬНОГО ФОРМАЛИЗМА СУЩНОСТЬ-СВЯЗЬ-АТТРИБУТЫ

Гл. ас. д-р инж. Арсов С., Докторант инж. Арсова Е.
Русенски университет „Ангел Кънчев” - Русе, България

Abstract: The paper presents an experimental investigation of the entity/ relationship/ attributes (E/R/A) semantic data model, which is evaluated by comparison with the traditional and popular entity/relationship (E/R) data model. The E/R/A modeling approach, a more powerful version of the original E/R model, is enriched with a new unconventional element, which represents relationships between entities and their own attributes. The new relationship is characterized with its semantics (name), type, the entity's name, and the name of each entity's attribute and enables a user to obtain a direct access to all data elements by queries written in a restricted natural language.

KEYWORDS: DATA MODEL, ENTITY-RELATIONSHIP-ATTRIBUTES, ENTITY-RELATIONSHIP, MODELING FORMALISM

1. Увод

Моделът на данните е интелектуално средство, което се използва при моделиране на част от реалния свят, интересуваша дадена организация. В качеството си на средство за моделиране той осигурява множество от понятия, които могат да бъдат използвани за определяне на присъщата структура на данните от реалния свят и на операциите, които са разрешени за изпълнение над тях. В качеството си на средство за изобразяване моделът на данните предава на потребителите, аналитиците и разработчиците на информационни системи представата на проектанта за данните.

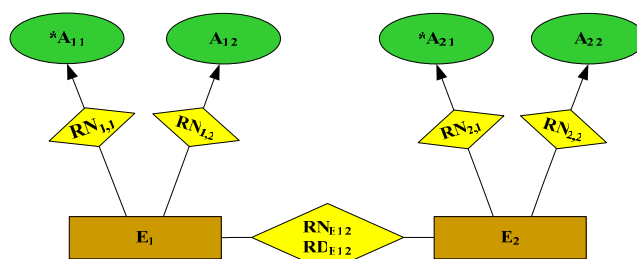
За да бъдат приемливи и полезни за крайните потребители, моделите на данните трябва да бъдат изследвани от гледна точка на човешкия фактор по отношение на разбираемост, използваемост и полезност [3]. Няколко предшествващи проучвания са изследвали ефектите от различни формализми за моделиране на данни [2], [3], [4]. Те се различават по четири размерности: участващите субекти; сравняваните модели на данните; експерименталните задачи; зависимите критерии.

Основна цел на представеното експериментално изследване е да докаже работоспособността и да изследва ефикасността на разглеждания в [1] моделен формализъм Същност-Връзка-Атрибути (С/В/А), както и ефективността на дейността на системните потребители при неговото използване. Методиката за провеждане на експерименталното изследване на модела С/В/А е обединение на методиките, използвани в средни съвременни изследвания. Това дава възможност като резултати от експеримента да бъдат получени повече различни оценки.

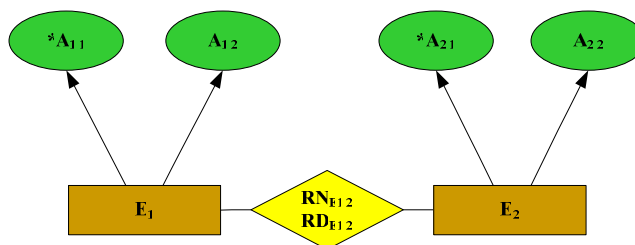
2. Разработване на методика за експериментално изследване

2.1. Формулиране на изследователските въпроси

Основната разлика между моделните формализми С/В/А и „Същност-връзка” (С/В) е методът, по който те представят връзките между елементите на данните. Синтаксисът на модела С/В/А е представен на фигура 1., а синтаксисът на модела С/В на фигура 2. Във връзка с постигане на основната цел на експерименталното изследване са формулирани следните три типа въпроси:



Фиг. 1. Елементи от диаграма "Същност-връзка-атрибути"



Фиг. 2. Елементи от диаграма „Същност-връзка”

Първият тип въпроси, групата от въпрос № 1 до въпрос № 5, се отнасят до полезността на алтернативните моделни формализми за процеса на обосновка, а именно: 1) Влияят ли различията на моделния формализъм С/В/А, от гледна точка на представянето на данните, върху способността на системните потребители да разбират базисната действеност? (обобщен въпрос); 2) Пораждат ли различията на модела С/В/А, от гледна точка на представянето на данните, различни образци при тълкуването или разбирането? (обобщен въпрос); 3) Дали предложеният модел С/В/А представя и предава повече семантика на данните към потребителите, отколкото моделът С/В? (въпроси 4, 5, 6 и 7); 4) По-лесен ли е за обосноваване (validation), включващо разбиране и проверка за противоречивост, моделът С/В/А от модела С/В? (въпроси 8, 9, 10, 11 и 12); 5) По-разбираеми ли са за системните потребители понятията, използвани в модела С/В/А от модела С/В? *Вторият тип въпроси*, които са обобщени във въпрос № 6, се отнасят до влиянието на модела С/В/А върху ефективността на дейността на системните потребители при формулиране на въпроси към базите от данни (БД), а именно: 6) Влияят ли различията на моделите на данните върху способността на

системните потребители да формулират въпроси на естествен език, базирани се на концептуални схеми (КС), проектирани чрез използване на различните формализми за моделиране? *Третият тип въпроси* са обобщени във въпрос № 7, а именно: 7) Какво е субективното възприемане на степента на трудност при използване и полезността на моделния формализъм? Този обобщен въпрос се детайлизира чрез група от шест въпроса.

2.2. Дефиниране на модела на изследването

Във връзка с доказване повишаването ефективността на дейността на потребителите при използване на нововъведения от автора модел С/В/А и при формулиране на въпроси към БД, се поставя задачата за изследване и сравняване броя на правилно формулираните въпроси на естествен език. Въпросите се формулират, като се използват КС на БД, които се представят съответно чрез моделите С/В/А и С/В. При определяне на информационните изисквания, както е описано в [3] и [4], един от етапите, изпълнявани от потребителя, е обосновката на модела. Преди да се направи заключение, че моделът е коректен, логичен и завършен, той трябва да бъде обоснован. Ето защо единият от аспектите на изследването е да се оцени въздействието на моделния формализъм върху производителността на потребителя при обосновката на модела. Двата аспекта на обосновката са: разбиране и проверка на противоречията. Потребителят трябва да разбере смисъла на модела, като след това той трябва да идентифицира противоречията между модела и неговите знания от действителността. Като допълнение на обективните измервания на производителността се събират данни за една важна *променлива*, възприемана полезност, посредством интервюиране чрез въпросник. Чрез тази променлива се измерва лекотата на използване на диаграми на БД, разработени чрез употреба на съответния моделен формализъм така, както тя се възприема от субектите.

• *Дефиниране на независимите променливи (фактори):* Всички външни въздействия върху изследвания обект в рамките на експеримента се наричат независими променливи или фактори. При провеждания от авторите експеримент едната от независимите променливи е видът на моделния формализъм (С/В/А и С/В). Втората независима променлива е типът на решаваната задача. При провеждането на експеримента всеки субект се отнася към една от двете групи по случаен принцип и се обучава в съответния моделен формализъм;

• *Дефиниране на зависимите променливи (параметри):* Величините, чрез които се описват реакциите на изпитвания обект, породени от външните въздействия, се наричат зависимите променливи, параметри или отклици. На базата на задачите, поставени в подраздел 2.1. са дефинирани три зависимите променливи (параметри): 1) производителност на субектите при обосновка на модела; 2) относителен дял на коректно

формулирани въпроси и полезност на съответния моделен формализъм в това отношение; 3) субективно възприемана полезност на моделния формализъм.

Рамката за оценяване на производителността на потребителя при изпълнение на задачите за обосновка на модела има два основни компонента: синтактичен и семантичен. Във връзка с това производителността на потребителя при обосновка на модела се оценява по два критерия. Първият критерий включва разбирането на конструкциите на моделния формализъм от субектите, т.е. чрез него се оценява производителността на потребителите при разбиране на синтаксиса на елементите на БД. Вторият критерий е свързан с коректността при проверка на противоречивостта, т.е. чрез него се оценява производителността на потребителите при разбиране семантиката на представените чрез модела БД. Производителността при разбиране на конструкциите на моделния формализъм се измерва чрез броя на коректните отговори на въпроси, отнасящи се до основните моделни конструкции. Оценката се базира на схемата, която е разработена от Juhn и Naumann [3] и е използвана от Kim и March [4];

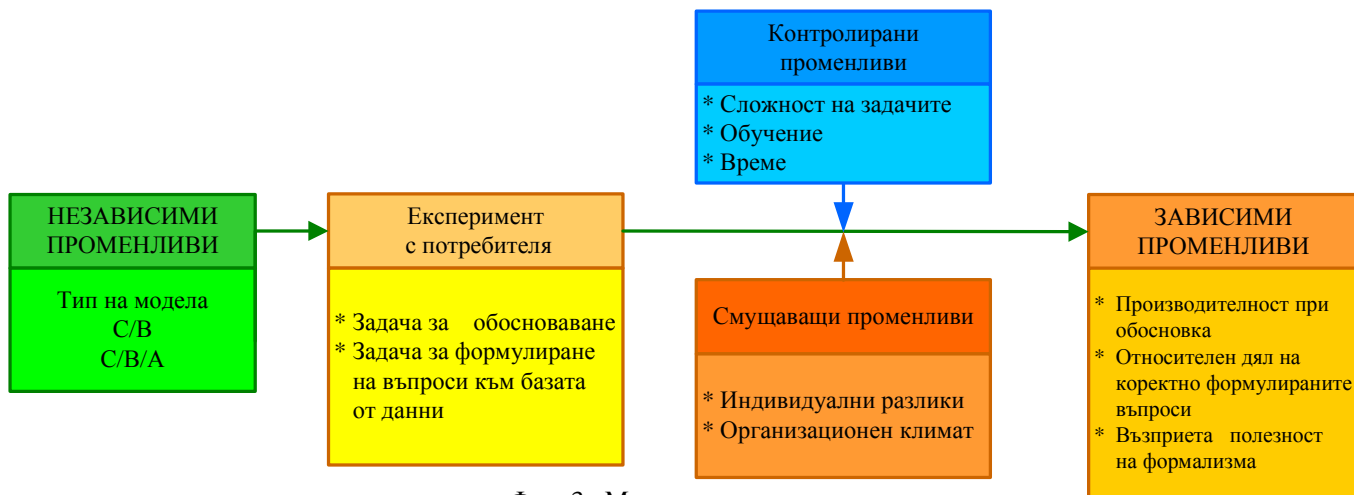
• *Определяне на контролираните променливи:* За вземане на предпазни мерки против смущаващите ефекти, три променливи се контролират по време на експеримента: степен на обучение, време за решаване и сложност на задачите;

• *Дефиниране модела на изследването:* Експерименталните задачи включват обосновка (validation) на модела и преценка на качествата на модела от различни гледни точки, като общ критерий е “качество на резултатите”. Моделът на изследването изобразява връзките между променливите (фактора и параметрите), задачите и субектите на изследването. Моделът на изследването е дефиниран въз основа на посочените основни задачи и е представен на фигура 3.

2.3. Експериментален инструмент

• *Избор на субектите, участващи в експерименталните изследвания:* В този експеримент участват студенти в качеството на потребители на системата. При провеждане на контролиран лабораторен експеримент се изследва ефектът от тези характеристики на моделните формализми, които имат отношение към изобразяването на реалния свят, върху потребителското разбиране и използване на отделни БД;

• *Разработка на модел за представяне на данните:* За експериментално изследване са избрани моделни формализми С/В/А и С/В. Те са различни по отношение на ударението, което поставят върху представяните от тях базисни конструкции, същности, връзки между същностите, атрибути и връзки между същностите и собствените им атрибути. Основната разлика между тях обаче е начинът, по който те



Фиг. 3. Модел на изследването

представят връзките между елементите на данните: същности и собствените им атрибути. Представянето на връзките се реализира в три измерения: наличието на връзката; степента на връзката; името на връзката.

Експерименталният пример трябва да се базира на обичайната работна среда на субектите [3]. Поради това за решаване на експерименталните задачи, е избран типичен пример на предметна област-факултетна канцелария в университет. На всеки студент, участващ в експеримента, се предоставя проверена коректна КС на БД, проектирана чрез използване на един от изпитваните моделни формализми: С/В/А и С/В. Моделът С/В/А на експерименталната БД, описваща частично университетска канцелария, се представя чрез диаграма. Тя съдържа 8 същности, 10 връзки между същностите, 21 атрибута и 21 връзки между същностите и собствените им атрибути.

2.4. Разработка на въпроси и задачи за решаване от потребителите по време на експеримента

- **Въпроси и задачи за оценка на производителността на потребителите при обосновка на модела:** Дефинират се три типа задачи, които характеризират процеса на обосновка. Всяка от тези експериментални задачи се извлича от изследователските въпроси, формулирани в подраздел 2.1. Общият брой на дефинираните задачи за измерване на производителността е петнадесет. Групите от въпроси, зададени на участниците в експеримента, са следните: 1) Въпроси от общ характер; 2) Въпроси за разкриване семантичните връзки между същностите и същностите и собствените им атрибути; 3) Задачи за идентификация степените на връзките между елементите на данните; 4) Въпроси за разбиране на идентификатори. Във връзка с посочените въпроси е предложена скала за оценка на производителността при обосновка на модела;

- **Въпроси и задачи за измерване влиянието на моделния формализъм С/В/А върху потребителите за повишаване на производителността им при задаване на коректни въпроси на ограничен естествен език (ОЕЕ) към БД при предоставена им КС на БД:** На всеки субект, участващ в изследването, се предоставя КС, проектирана чрез използване на един от моделните формализми С/В/А или С/В. На всеки субект се поставя задача, като използва само един вид КС, да формулира десет въпроса към съответната БД. След формулиране на въпросите от страна на субектите се отчита броят на коректно формулираните въпроси. Чрез съпоставяне на броя им с общия брой въпроси се отчита относителния дял на коректните въпроси, зададени чрез използване на съответния формализъм за моделиране;

- **Въпроси и задачи за оценка на субективно възприеманата от потребителите полезност на моделите С/В/А и С/В:** Като допълнение на обективните измервания на производителността да се съберат данни от субектите чрез въпросник за една важна променлива – възприемана полезност. Чрез тази променлива се измерва лекотатата на използване и полезността на моделния формализъм така, както тя се възприема от субектите;

- **Хипотези, отнасящи се към формулираните въпроси и задачи:** Във връзка с разработените въпроси и задачи са формулирани следните хипотези: Х₁) Субектите, използващи семантичния модел С/В/А, ще намерят повече смислови връзки между елементите на данните, отколкото тези, използващи С/В; Х₂) Субектите, използващи семантичния модел С/В/А, ще идентифицират по-точно степените на връзките между елементите на данните, отколкото тези, използващи С/В; Х₃) Няма да има значителни разлики между групите, прилагащи различните моделни формализми, при разбирането

от тяхна страна на идентификаторите на елементите на БД; Х₄) Ще има значителни разлики между групите, прилагащи различните моделни формализми, по отношение на броя на правилно формулираните от тях въпроси към БД при зададена КС на БД.

2.5. Метод на дисперсионния анализ (*Analysis Of Variance (ANOVA)*)

При настоящия експеримент се изследва влиянието на един качествен (неизмерим) фактор (независима променлива), какъвто е факторът моделен формализъм, дефиниран в подраздел 2.2, върху параметрите (зависимите променливи), дефинирани в същия подраздел. В статистическите изследвания такива задачи се решават чрез разработения от Р. Фишер метод на дисперсионния анализ (ANOVA). В представеното експериментално изследване се използва методът на еднофакторния дисперсионен анализ.

3. Резултати и дискусия

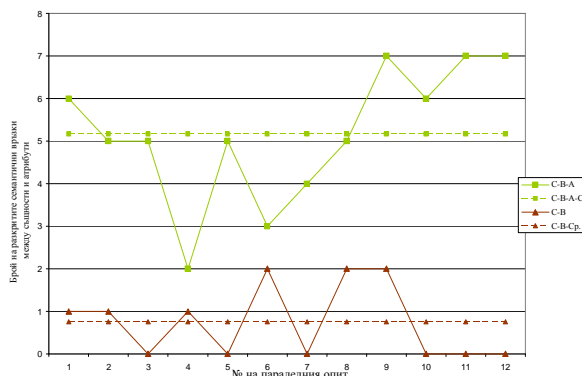
3.1. Влияние на моделните формализми С/В/А и С/В върху производителността на потребителите при обосновка (validation) на КС на БД

- **Резултати от изпълнението на задачите за разкриване на семантични връзки между елементите на данните в БД въз основа на КС:** Обобщените резултати от еднофакторния дисперсионен експеримент с моделните формализми С/В/А и С/В за разкриване на семантични връзки между същностите и между същностите и собствените им атрибути в БД, въз основа на зададена КС на БД, са представени в табл. 1 и на фигура 4.

Таблица 1. Резултати от еднофакторния дисперсионен експеримент, за откриване на връзки в моделите С/В/А и С/В

№ на паралелния опит	1	2	3	4	5	6	7	8	9	10	11	12	$y_{i.} = \sum_{j=1}^6 y_{i,j}$	\bar{Y}_i	
Нива на фактора моделен формализъм	С/В/А	6	5	5	2	5	3	4	5	7	6	7	7	62	5,17
	С/В	1	1	0	1	0	2	0	2	0	0	0	9	0,75	

В таблицата на дисперсионния анализ (табл. 2) са представени статистически обработените резултати, получени от експеримента за разкриване на семантични връзки между същностите и собствените им атрибути в БД, въз основа на зададена КС на БД. На базата на дисперсионния анализ може да се направи изводът, че има значителна разлика между групите, прилагащи при експеримента новият моделен формализъм С/В/А и групите, прилагащи



Фиг. 4. Графично сравнение производителността на субектите при разкриване на семантични връзки между същности и атрибути

Таблица 2. Дисперсионен анализ на резултатите

Източник на разсейване	Суми от квадратите	Степени на свобода	Оценки на дисперсията	Критерий на Фишер
Моделен формализъм	$SS_M = 117,04$	$2 - 1 = 1$	$S_A^2 = 117,04$	$F_M = 71,8$
Случайни и неотчетени фактори	$SS_R = 35,92$	$2 \cdot 12 - 2 = 22$	$S_R^2 = 1,63$	
Сумарно влияние	$SS = 152,96$	$2 \cdot 12 - 1 = 23$	$S^2 = 6,65$	

моделния формализъм С/В от гледна точка на семантичните връзки, разкрити между една същност и собствените ѝ атрибути. Това заключение се базира на сравнението на изчислената стойност на критерия на Фишер - F_M , която е по-голяма от критичната стойност - $F_{0,05;1;22}$,

($F_M = 71,8 > F_{0,05;1;22} = 4,31$). Следователно разработеният моделен

формализъм С/В/А влияе по-съществено върху производителността на системните потребители при разкриване на семантичните връзки между същностите от БД и техните собствени атрибути, при зададена КС на БД, отколкото моделния формализъм С/В.

След сравнителен анализ на резултатите от експерименталните изследвания върху производителността на системните потребители при разкриването на семантични връзки между елементите на данните в БД може да се направи следния обобщен извод: Моделният формализъм С/В/А предава по-голямо количество семантична информация към потребителите благодарение на семантичната информация, която за разлика от моделния формализъм С/В допълнително се представя в КС на БД чрез наименованите връзки между същностите и собствените им атрибути;

- *Резултати от изпълнението на задачите за идентификация на степените на връзките в БД на основата на КС:* Въз основа на дисперсионния анализ на резултатите, получени при идентификация на степените на връзките между същности и степените на връзките между същности и собствените им атрибути, може да се направи следното заключение: За общото повишаване на производителността на потребителите при идентификация на степените на връзките между елементите на данните допринасят допълнителните семантични връзки (посочени чрез средствата на моделния формализъм С/В/А) между същностите и собствените им атрибути;

- *Резултати от изпълнението на задачите за разбиране на идентификатори в БД, на основата на КС:* Моделните формализми С/В/А и С/В нямат съществена разлика по отношение на влиянието им върху потребителите за разбиране на идентификаторите на елементите на БД.

3.2. *Влияние на моделните формализми С/В/А и С/В върху производителността на потребителите при задаване на коректни въпроси на ОЕЕ към БД*

При зададена диаграма С/В/А на БД броят на коректно формулираните въпроси на ОЕЕ от системните анализатори

към БД е съществено по-голям, отколкото при зададена диаграма С/В на БД. От това следва, че представянето на БД чрез модела С/В/А влияе съществено върху производителността и качеството на работата на системните анализатори при формулиране на въпроси на ОЕЕ към БД.

3.3. *Оценки на моделните формализми С/В/А и С/В, базирани на субективното мнение на участниците в експеримента*

Експерименталните резултати от субективните оценки за неудобствата на двата моделни формализма С/В/А и С/В, показват, че според потребителите в това отношение няма съществена разлика между тях. По отношение на полезността и леснината на формулиране на въпроси на ограничен естествен език, въз основа на модела С/В/А и модела С/В, по-високо се оценява моделният формализъм С/В/А.

4. Заключение

Резултатите от експерименталните изследвания доказват работоспособността на предложения моделен формализъм С/В/А. Моделът С/В/А предава по-голямо количество семантична информация към потребителите в сравнение с модела С/В. Представянето на реалния свят чрез модела С/В/А влияе положително върху производителността и качеството на работата на системните анализатори при формулиране на въпроси на ОЕЕ към БД.

5. Литература

1. Arsov, S., B. Rachev. An Approach for Designing a Restricted Bulgarian Natural Language Database Query System. - In: Proceedings of 19th International CODATA Conference "The Information Society: New Horizons for Science", Berlin, Germany, 7-10 November 2004. <http://www.codata.org/04conf/papers/Arsov-paper.pdf> (14.03.2010)
2. Batra, D., J. Hoffer, and R. Bostrom. A Comparison of User Performance Between the Relational and the Extended Entity Relationship Models in the Discovery Phase of Database Design. Communication of the ACM, Vol 33, № 2, February, 1990, pp.126-139.
3. Juhn, S., and J. Naumann. The Effectiveness of Data Representation Characteristics on User Validation. - In: Proceedings of the 6th International Conference on Information Systems, Indianapolis, Indiana, USA, 1985, pp.212-226.
4. Kim, Y., and S. March. Comparing Data Modeling Formalisms. Communications of the ACM, Vol. 38, № 6, June 1995, pp.103-115.

Изследванията са извършени/подпомогнати по Договор № BG051PO001/07/3.3-02/8 „Механизми за осигуряване качествено израстване на научните кадри”, финансиран по схема "Подкрепа за развитие на докторанти, постдокторанти, специализанти и млади учени" на ОП "Развитие на човешките ресурси" на "Европейския социален фонд".