РЕЦЕНЗИЯ

на дисертационен труд за присъждане на образователна и научна степен „Доктор”
в професионално направление 4.6 Информатика и компютърни науки,
научна специалност 01.01.12 Информатика
на тема **Приложение на Data Warehouse системи за разкриване на знания: Разреденост на данните в многомерния аналитичен модел**
с автор Ина Асенова Найденова, ФМИ, СУ „Климент Охридски”
от доц. д-р Каталина Петрова Григорова, Русенски университет „Ангел Кънчев”

1. Общо описание на дисертационния труд и на приложените към него материали

Дисертационният труд на Ина Найденова е посветен на актуален проблем, свързан с развитието на аналитични системи, които предоставят на съвременните компании възможности за многоаспектен анализ, прогнозиране и оптимизиране на дейността им.

Работата е с обем от 145 стр. и съдържа увод, 5 глави (15, 15, 10, 26, 33 страници съответно), заключение, перспективи за бъдещо развитие, анотация на получените резултати с формулирани научно-приложни и приложни приноси, списък на литературни източници (97 заглавия), списък с използваните съкращения на английски и 3 приложения (9, 12, 3 страници съответно).

В уводната част на дисертацията, освен актуалността на избрания проблем, е представена и цялостната методология на изследването.

Целта е дефинирана ясно – изследване на явлението разреденост на данните при многомерен анализ в складовете от данни, анализиране на проблемите, произтичащи от това явление, както и предлагане на ефективни решения на изследваните проблеми. Формулираните **изследователски задачи** съответстват на целта:

- задълбочено проучване и анализ на многомерния модел на данните, неговите предимства и ограничения в системите за разкриване на знания;
- систематично изследване на принципите и методите за преодоляване на ограниченията на многомерния модел, като се предложат подобрения и/или нововъведения към тях;
- формално описание на предложените подобрения и/или нововъведения;
- извършване на експерименти и тестове в подкрепа на направените изследвания и предложения.

Формулираните задачи подкрепят постигането на целта и очертават един завършен цикъл на изследването.

Глава 1 е посветена на многомерния модел на данни. Разгледани са характеристиките на складовете от данни, като източници за обобщени данни от оперативните и транзакционни системи на организациите, които се структурират по начин, удобен за анализ и извличане на справки за целите на фирменото управление. Обоснована е необходимостта от обработване на големи обеми информация чрез използване на складове от данни и средства за аналитична обработка. Подчертана е ролята на многомерния модел на данни в процеса на анализ. Детайлно са описани основните понятия в него и са показани начини за дефиниране на аналитични справки.

Съществено внимание е отделено на ограниченията на многомерния модел – разреденост и експлозия на данни, които затрудняват неговата реализация. Представени са архитектурни подходи за реализация на многомерния модел и са анализирани техните предимства и недостатъци.

Глава 2 представя методи за съхраняване и извличане на данни в аналитичните системи. Посочени са предимствата на съхраняване на данни в многомерен модел в сравнение с релационния по отношение на заемано дисково пространство. Анализирани са възможности за ограничаване на експлозията на данни чрез използване на техники за ефективно съхраняване на моделите. Разгледани са структури от данни, използвани за съхраняване на данните и улесняване на достъпа до тях, както и за намаляване на времето за отговор при изпълнение на заявки за извличане на данни.

В **Глава 3** се разглежда класификация на свойството разреденост на данните. Посочени са видове разреденост, коментирани в различни литературни източници. Въз основа на наблюдение, че разредеността е следствие на наличието на различни ограничения, е създадена нова класификация на базата на семантиката на разредеността. Дадени са дефиниции на видове разреденост. Описани са видове семантични връзки между стойностите на измеренията, които водят до регулярна разреденост и са предложени примери, поясняващи видовете връзки. Демонстрирани са подходи за намаляване на разредеността.

В **Глава 4** е предложена формална дефиниция на многомерния модел на данни – концептуален кубов модел. Дефиницията е направена след разглеждане на различни дефиниции на този модел, описани в литературните източници. Изборът на предложената дефиниция е обоснован със стремежа да се използва представяне, което е възможно най-близко до широко разпространената представа за модела като съвкупност от кубове в *n*-мерното пространство.

Предложено е разширение на многомерния модел на данните – карта на регулярна разреденост, използвано за различаване на регулярната от случайната разреденост. Описани са подходи за представяне на областите на регулярна разреденост, като специално внимание е отделено на представяне на картата с използването на множество от правила в „атомарен” вид. Всяко правило определя област на регулярна разреденост в куба, а обединението на всички правила определя цялата карта на регулярна разреденост. Разгледани са възможни разширения на приетото представяне, предоставящи по-голямо удобство на проектантите на многомерния модел при създаване на карти на регулярна разреденост. Демонстрирани са начини за преобразуване на бизнес правила до избрания „атомарен” вид.

Анализирани са възможни приложения на картата на регулярна разреденост:

- за подобряване на качеството на данните и подпомагане на процеса на откриване на грешки;
- за повишаване на производителността и намаляване на заеманото пространство;
- за интерпретация на данните в аналитичните системи.

Описан е метод за откриване на области с регулярна разреденост, използван за реализацията на редактор за конструиране на карти на регулярна разреденост и извличане на данни от нея. Направена е обосновка на сходимостта на метода.

Глава 5 представя приложения на направените изследвания. Разгледани са експериментални резултати в три основни направления:

- отстраняване на повторения на данни в адитивни йерархични участъци на неадитивни показатели;
- сравнителен анализ на заеманата памет и бързодействието при елиминирането на връзки от тип ирелевантност и сегментация на измерения;
- ефективност от прилагането на подхода на сливане на измерения.

Експериментите са извършени като е използван реално съществуващ склад за данни, съхраняващ информация за международен холдинг.

Описан е и редактор за създаване на карта на регулярна разреденост и реализиране на метода за откриване на области с регулярна разреденост.

Експерименталните резултати са много добре и подробно онагледени с богат графичен материал – таблици, графики, екрани.

2. Актуалност на проблема

Актуалността на разглеждания в дисертацията проблем е обоснована обширно в уводната част на дисертацията – нарастването на обемите от данни, използването на комплексни и все по-сложни модели на данните, необходимостта от гъвкави и динамични аналитични решения, които дават както единен корпоративен поглед върху дадена компания, така и различни аналитични решения, съобразени със спецификата на нейните отделни подразделения.

Натрупването на големи обеми от данни, резултат от напредъка на технологиите за събиране и съхраняване на бизнес данни, е предпоставка за развитие на системи за интерактивни и многоаспектни анализи, които позволяват на крайните потребители да формират и проверяват различни хипотези, да прогнозират и оптимизират дейността на компаниите.

Многомерният модел на данни, който е в основата на системите за бизнес анализ, е обект на изучаване и развитие с цел удовлетворяване на повишените очаквания на потребителите за следене и анализ на бизнеса.

Настоящият дисертационен труд, фокусиран върху някои ограничения на многомерния модел, допринася за неговото изучаване и развитие.

3. Познаване състоянието на проблема

Ина Найденова демонстрира задълбочено познаване на състоянието на проблемите по темата на дисертацията, както чрез обхвата, така и чрез дълбочината на интерпретиране на използваните литературни източници.

Списъкът на използваната литература включва 97 референции, от които 15 от Интернет адреси. По-голяма част от източниците са публикувани след 2000 година – 84, 5 са на български език, а останалите – на английски.

От съдържанието на глава 1 и глава 2 може да се заключи, че докторантката е добре запозната с историята и текущото състояние, както на многомерния модел на данните, така и на системите за бизнес анализ и методите за ефективно съхранение и извличане на данни.

Анализът на литературата, на особеностите на многомерния модел, на методите за съхраняване и извличане на данни и спецификите на свойството разреденост може да се приеме за съществен принос.

Насоките на предложените в следващите глави изследвания са обосновани и са следствие на направения задълбочен анализ.

4. Подход и решение на проблема

Методологическите параметри на изследването са конкретизирани в увода. Дефинирани са целите и задачите на дисертационния труд. Решението на проблема е представено в глави 2, 3 и 4.

Аналитичният обзор на въпроси, свързани с оптимизацията на заеманото физическо пространство и производителността на заявките в системите за бизнес анализ, води до заключенията, че съществуващите методи за преодоляване на недостатъците на многомерния модел не отчитат природата на разредеността в модела и не използват знания за семантиката и бизнес ограниченията от съответната предметна област.

Тези заключения се явяват предпоставка за детайлното разглеждане на свойствата на явлението разреденост и причините за наличието му. Показано е как познаването на връзките между стойности на отделни измерения може да доведе до намаляване на разредеността.

Като резултат е предложена нова класификация на видовете разреденост, която се използва при разработването на подходи за намаляване на регулярната разреденост. Представената в глава 4 карта на регулярна разреденост е оригинален обект, който се явява разширение на многомерния модел и способства за подобряване на качеството на данните, намаляване на заеманата памет и увеличаване на производителността на обработката на заявки.

Вижда се, че авторката притежава необходимата научна култура за прилагане на резултатите от системното проучване на текущото състояние на разглеждания проблем. Прецизно и подробно са описани предложените допълнения и разширения на многомерния модел на данните, както и резултатите от тяхното приложение на примера на реално съществуващ склад от данни.

Важно е да се отбележи и много прегледното и грижливо оформление на дисертацията.

5. Основни приноси

Вече отбелязах, че анализът на ограниченията на многомерния модел на данни и на методите за ефективно съхранение и извличане на данни в аналитичните системи е принос на докторантката, който може да бъде класифициран като научно приложен;

Към тази група могат да бъдат отнесени и:

- въведената нова класификация на видовете разреденост в *n*-мерен куб;
- систематизираните видове семантични връзки между стойностите на измеренията в даден *n*-мерен куб, водещи до появата на регулярна разреденост;
- представеният нов подход за намаляване на регулярната разреденост, разширението на многомерния модел на данните чрез въвеждане на оригинален обект – карта на регулярна разреденост;
- изследването на приложимостта на картата на регулярна разреденост за подобряване на коректността на данните, ефективното им извличане, съхранение и интерпретация;
- формализираният метод за откриване на области с регулярната разреденост в *n*-мерен куб и формализираното представяне на карта на регулярна разреденост като съвкупност от бизнес правила;
- методът за отстраняване на повторения на данни в адитивни йерархични участъци на неадитивни показатели.

Приложните приноси са свързани с изследване на ефективността на методи за подобряване на качеството на данните и проектиране на софтуерна среда, реализираща предложеното представяне и някои от приложенията на картата на регулярна разреденост.

6. Публикации по темата на дисертацията и личен принос на автора

Резултатите от дисертационния труд са представени в 5 публикации, които са на английски език. Те засягат основно резултатите от глави 3, 4 и 5, в които авторката излага своите идеи за решаване на разглеждания проблем.

Публикациите могат да бъдат класифицирани по няколко признака:

- 2 от публикациите са самостоятелни и 3 в съавторство;
- 3 са публикувани в България, 1 в Португалия и 1 в Румъния;

- 2 публикации са в международни научни списания и 3 в сборници от международни конференции.

Открити са и две цитирания на две от публикациите.

Смятам, че една публикация, свързана с направения аналитичен обзор на особеностите на многомерния модел и на свойството разреденост би допринесла за още по-пълното покриване на работата на докторантката.

Общата ми преценка за публикационната дейност на Ина Найденова е положителна – дисертационният труд е покрит достатъчно и е получил известност сред научната общественост в чужбина и у нас.

Отчитайки процента на самостоятелните авторски публикации по темата, считам, че всички разработки, които са намерили място в текста, са лично дело на Ина Найденова, подкрепено от нейния научен ръководител и съобразено с идейните предпоставки, залегнали в съществуващи подобни реализации.

7. Автореферат

Авторефератът, в размер от 36 страници, отразява съществените аспекти на дисертационния труд. Постигнато е балансирано покритие на основните етапи и приноси. Освен че е добре структуриран, авторефератът включва и списъци на публикациите по дисертацията и на известните цитирания.

8. Критични бележки и препоръки

В текста се срещат две понятия „карта на регулярна разреденост" и „карта на регулярна рехавост", за които се предполага, че са идентични. За предпочитане е да се използва само едното от тях.

Добре е в края на глава 5. също да бъдат формулирани приноси и заключение.

Малко повече примери за тестване на редактора на карта за регулярна разреденост биха способствали за повече изводи и обобщения на резултатите от тестването.

Литературните източници в автореферата и в дисертацията се различават по брой. Би следвало да има пълно съответствие на библиографията в двата текста.

В библиографията на дисертацията са включени заглавия, които не са цитирани в текста.

Макар и рядко, срещат се някои правописни и езикови неточности: например на стр. 54 – „последяващ анализ" вместо „анализ, проследяващ…"

Библиографията не е съвсем по азбучен ред.

8. Заключение

Ина Найденова е представила цялостно и завършено изследване с добра логическа структура, което напълно покрива изискванията за присъждане на образователната и научна степен „доктор". Дисертационният труд отговаря на изискванията на ЗРАС. Давам му обща положителна оценка.

Предлагам на Ина Найденова да бъде присъдена образователната и научна степен „доктор" в професионално направление 4.6 Информатика и компютърни науки, научна специалност Информатика.

03.12.2012 г. Рецензент:

/доц. д-р Каталина Григорова/