Екатерина Маградзе

Разработка и исследование моделей оценок системы налогового контроля

Представлена на соискание академической степени
доктора наук

Грузинский Технический Университет
Тбилиси, Грузия
Октябрь, 2008г

საქართველოს ტექნიკური უნივერსიტეტი

ეკონომიკური ინფორმატიკა

ჩვენ, ქვემორე ხელისმომწერნი ვადასტურებთ, რომ გავეცანით ეკატერინე მაღრაძის მიერ შესრულებულ სადისერტაციო ნაშრომს დასახელებით: “Разработка и исследование моделей оценок системы налогового контроля” (“საგადასახადო სისტემის მუშაობის შეფასების მოდელების დამუშავება და კვლევა”) და ვაძლევთ რეკომენდაციას საქარველოს ტექნიკური უნივერსიტეტის ეკონომიკური ინფორმატიკის სადისერტაციო საბჭოში მის განხილვას დოქტორის აკადემიური ხარისხის მოსაპოვებლად.

ხელმძღვანელი:	პროფ. ტ.მ.დ. ზურაბ ბოსიკაშვილი
რეცენზენტი:	პროფ. ტ.მ.დ. რომან სამხარაძე
რეცენზენტი:	პროფ. ტ.მ.დ. ავთანდილ კვიტაშვილი
რეცენზენტი:	

საქართველოს ტექნიკური უნივერსიტეტი

2008 წელი

ავტორი: მაღრაძე ეკატერინე

დასახელება: “საგადასახადო სისტემის მუშაობის შეფასების მოდელების დამუშავება და კვლევა”

ფაკულტეტი : ეკონომიკური ინფორმატიკა

ხარისხი: დოქტორი

სხდომა ჩატარდა:

Оглавление

.

Резюме

Налоговая служба начиная с 2007 года отказалась от тотального контроля всех налогоплательщиков. Т.е. больше не ставится цель проверять всех поголовно. Теперь ревизоры делят налогоплательщиков на тех, которые требуют повышенного внимания, и тех, кого достаточно «мониторить» по имеющимся в представленных декларациях данным. Поэтому, сегодня выход на выездную проверку – это результат уже продуманных и выверенных действий налоговой инспекции с определением конкретных зон риска у налогоплательщика. Для этого и была утверждена система планирования выездных налоговых проверок. Целью такого планирования является концентрация контрольных мероприятий на зонах риска, обеспечение качества проверок и мотивирование налогоплательщиков к добровольному отказу от инструментов минимизирования налогов.

Утвердив эту систему, мы сделали процесс планирования налоговых проверок открытым. Разработали предложения по совершенствованию организации и методов контрольной работы налоговых органов и определили критерии оценки ее эффективности с помощью использования новых технологий таких как Data Mining, Хранилища Данных.

Используя технологии извлечения знаний, включая постановку задачи, подготовку данных, автоматическое построение моделей, анализ и тестирование результатов, использование моделей в реальных приложениях мы достигли желаемых результатов. Теперь достаточно обратиться к программе оценки моделей и за считанные минуты выявить : наиболее вероятные «зоны риска» (нарушения законодательства о налогах и сборах), своевременно отреагировать на возможное совершение налоговых правонарушений и определить необходимые мероприятия налогового контроля.

Целью диссертационной работы является обоснование системного построения налогового контроля и разработка методического инструментария налоговых расследований для противодействия налоговым нарушениям.

Достижение поставленной цели потребовало решения следующих задач:

- Построение и усовершенствование информационно-аналитической системы Министерства Финансов Грузии;
- Создание хранилищ данных, что позволяет существенно повысить общую эффективность создаваемой информационной системы. Хранилище данных может объединять информацию из текстовых файлов и многих баз данных, как реляционных , так и нереляционных, в единую систему поддержки принятия решений;
- Повышения скорости выполнения запросов с помощью создания материализованных представлений, как мощного средства повышения производительности хранилищ данных и систем поддержки принятия решений.
- Фрагментации таблиц как средства повышения масштабируемости при увеличении размеров больших объектов в базе данных, что положительно сказывается на производительности, доступности данных и упрощает администрирование.
- Организации процесса Data Mining, включающие, следующие фазы**:**

    1. осмысление бизнеса,

2. осмысление данных,
3. подготовка данных,
4. моделирование,
5. оценка результатов ,
6. внедрение.

- Систематизировать и предложить новые методы выявления недисциплинированных налогоплательщиков;

Научная новизна диссертационной работы заключается в исследовании и создании моделей, помогающих отыскивать скрытые зависимости в большом обьеме обрабатываемой информации. На данный момент применение Data Mining технологий на мировом рынке довольно распространено (разработкой Data Mining технологий занимаются такие крупные компании как ORACLE, IBM, MICROSOFT), однако относительно к развивающейся инфраструктуре налоговой системы Грузии освоение данной техлогогии стало новым этапом развития информационно-аналитической системы в целом, тем самым обеспечивая широкий спектр возможностей информационной деятельности, предоставляя средства для анализа ситуаций, принятия решений и реализации служебных полномочий и интеллектуальных потребностей налоговой службы.

Разбивая налогоплательшиков при помощи инструментов Data Mining на различные группы, мы имеем возможность сделать налоговую политику более целенаправленной, а потому - эффективной. Теперь мы не тратим время на выявление недобросовестных налогоплательшиков из многотысячного списка, опираясь только на опыт аудиторов. Создание моделей, на основании которых мы можем проводить качественный анализ данных, происходит в центральном оффисе. С помощью квалифицированных специалистов-аналитиков, путем изменения коэффициентов , с последующим пересчетом данных посредством хранимых проседур находится модель , наиболее точно отражающая действительность. При этом следует отметить, что в отличие от других методов поддержки принятия решений технологии Data Mining обладают гораздо более высокой степенью интеллектуальности, позволяя в значительной степени автоматизировать анализ данных.

რეზიუმე

2007 წლიდან საგადასახადო სამსახურმა უარი თქვა ყველა გადამხდელის კონტროლზე, ანუ აღარ დგას თითოეული მათგანის შემოწმების ამოცანა. ამჟამად გადამხდელების ერთი ნაწილი მოითხოვს განსაკუთრებულ ყურადღებას და აუცილებელ შემოწმებას, მაშინ როდესაც გადამხდელთა მეორე ნაწილზე დაკვირვება მხოლოდ წარდგენილი დეკლარაციების ანალიზით შემოიფარგლება. აქედან გამომდინარე, გასვლითი შემოწმება წარმოადგენს საგადასახადო სამსახურის კარგად მოფიქრებულ და შეჯერებულ მოქმედებას გადამხდელის კონკრეტული რისკის ზონის წინასწარი განსაზღვრით. ამიტომაც იქნა დამტკიცებული გასვლითი საგადასახადო შემოწმებების დაგეგმვის კონცეფცია . დაგეგმვის ძირითადი მიზანია, მოხდეს საკონტროლო ღონისძიებების კონცენტრაცია რისკის ზონაზე, შემოწმების ხარისხის უზრუნველყოფა და გადამხდელთა მოტივირება ნებაყოფლობით თქვან უარი გადასახადთა მინიმიზირების "ინსტრუმენტთა" გამოყენებაზე.

კონცეფციის დამტკიცებით, საგადასახადო შემოწმებათა დაგეგმვის პროცესი გამჭვირვალე გახდა, შემუშავებული იქნა წინადადებები საგადასახადო ორგანოების საკონტროლო ღონისძიებების ორგანიზების და მეთოდების გაუმჯობესების მიზნით, განსაზღვრულ იქნა ამ ღინისძიებების ეფექტურობის შეფასების კრიტერიუმები ისეთი ახალი ტექნოლოგიების გამოყენებით, როგორიცაა Data Mining და მონაცემთა საცავები.

ამოცანის დასმის, მონაცემთა მომზადების, მოდელების ავტომატური აგების, ცოდნის მოპოვების, შედეგების ანალიზისა და ტესტირების ტექნოლოგიების გამოყენებით, მათი რეალურ პროგრამებში მორგებით მიღწეულ იქნა სასურველი შედეგები. ახლა საკმარისია მივმართოთ პროგრამას და რამოდენიმე წუთში:

გამოვლენილი იქნება შესაძლო "რისკის ზონები" (საგადასახადო კანონმდებლობის დარღვევები).

მოხდება აუცილებელი საგადასახადო კონტოლის ღონისძიებების განსაზღვრა და საგადასახადო დარღვევებზე რეაგირება.

სადისერტაციო ნაშრომის მიზანს წარმოადგენს დაასაბუთოს საგასადახადო კონტროლის სისტემური აგების აუცილებლობა და საგადასახადო გამოკვლევების მეთოდოლოგიური საშუალებების შექმნა საგადასახადო დარღვევების აღმოსაფხვრელად. დასახაული მიზნის მისაღწევად შემდეგი ამოცანების გადაჭრა გახდა საჭირო:

მონაცემთა საცავის შექმნა, რომელიც არსებითად ამაღლებს საინფორმაციო სისტემის ეფექტურობას. მონაცემთა საცავი შეიძლება მოიცავდეს როგორც ტექსტურ ფაილებს, ასევე რელაციურ და არარელაციურ მონაცემთა ბაზებს. გადაწყვეტილების მიღების სისტემა დაფუძნებული იქნება მონაცემთა საცავზე.

- მონაცემთა ბაზებისადმი შეკითხვების შესრულების დაჩქარება მატერიალიზებული წარმოდგენების საშუალებით. ეს მიდგომა

გაცილებით ასწრაფებს მონაცემთა საცავის და გადაწყვეტილებების მიღების სისტემის მუშაობის წარმადობას.

- ცხრილების ფრაგმენტაცია, რომელიც საუკეთესო საშუალება ა დიდი ზომის მონაცემთა ბაზებში ინფორმაციის შესანახად, ეს კი დადებითად აისახება წარმადობასა და მონაცემთა წვდომაზე. შესაბამისად ადვილდება ადმინისტრირებაც.
- Data Mining-ის პროცესის ორგანიზება, რომელიც თავის მხრივ შემდეგ ფაზებს მოიცავს: ბიზნესის გააზრება, საჭირო მონაცემების მოპოვება და მათი მომზადება, შედეგების შეფასების მოდელირება და დანერგვა.
- არაკეთილსინდისიერი გადამხდელების გამოსავლენად ახალი მეთოდების შეთავაზება და სისტემატიზირება.

სადისტერტაციო ნაშრომის სამეცნიერო სიახლე იმ მოდელების შემუშავებასა და გამოკვლევაში მდგომარეობს, რომელთა დახმარებითაც შესაძლებელია დასამუშავებელ მონაცემთა დიდ რაოდენობაში ფარული დამოკიდებულებების აღმოჩენა. ამ დროისათვის Data Mining ტექნოლოგიის გამოყენება მთელს მსოფლიოში ფართოდაა გავრცელებული (ისეთი დიდი კომპანიები, როგორიცაა ORACLE, IBM, MICROSOFT ამ ტექნოლოგიის განვითარებაზე ზრუნავენ), მაგრამ საქართველოს განვითარებადი საგადასახადო სისტემის ინფრასტრუქტურისათვის მოცემული ტექნოლოგიის დაუფლება ზოგადად ინფორმაციულ–ანალიტიკური სისტემების განვითარების ახალ ეტაპს წარმოადგენს: დამუშავდება ინფორმაციის გაცილებით ფართო სპექტრი, გაჩნდება სიტაუციების ანალიზის, გადაწყვეტილებების მიღების და საგადასახადო სისტემის ინტელექტუალური მოთხოვნების დაკმაყოფილების და სამსახურეობრივი უფლებამოსილების უკეთ განხორციელების საშუალება.

Data Mining–ის ინსტრუმენტების გამოყენებით გადამხდელთა დაყოფა სხვადასხვა ჯგუფებად, საშუალებას გვაძლევს საგადასახადო პოლიტიკა უფრო მიზანმიმართული, და შესაბამისად, უფრო ეფექტურიც გავხადოთ. აღარ იქნება საჭირო მხოლოდ აუდიტორების გამოცდილებაზე დაყრდნობით არაკეთილსინისიერი გადამხდელების გამოსავლენა მრავალათასიანი სიიდან–შესაბამისად აღარ დაიკარგება დრო. ცენტრალურ ოფისში განხორციელდება მოდელების შექმნა, რომელთა საფუძველზეც ჩატარდება მონაცემთა ხარისხობრივი ანალიზი. რამოდენიმე კვალიფიციური სპეციალისტ–ანალიტიკოსი კოეფიციენტების შეცვლის გზით და შენახული პროცედურების გამოყენებით იპოვიან იმ მოდელს, რომელიც ყველაზე მეტად შეესაბამება სინამდვილეს. ამასთან აღსანიშნავია, რომ გადაწყვეტილებების მიღების სხვა მეთოდებისგან განსხვავებით, Data Mining ტექნოლოგია გაცილებით მეტი და უკეთესი ხარისხის ინტელექტუალური და გაფართოვებადი საშუალებებით გამოირჩევა და მონაცემთა ავტომატიზირებული ანალიზის საშუალებას იძლევა.

Rezume

The State Tax Service has refused the total control of all taxpayers since 2007. It is not a goal to check all of them. Now auditors divide taxpayers into different groups. The One group needs a special attention, and for another one it is enough to monitor the data available in presented declarations. Therefore, today the cameral audit is a result of thought over and verified actions of the tax inspection with definition of concrete zones of risk at the taxpayer. That is why the creation of planning system for cameral audits has been approved. The purpose of such planning is concentration of control actions on risk zones, maintenance of quality of checks and motivation of taxpayers to voluntary refuse a gap for minimizations of taxes. After implementation of this system tax audits planning process became transparent. There are developed offers for improving examination methods and organization of audits by tax service bodies, are defined criteria to estimate their efficiency using new technologies such as Data Mining, Datawarehouses.

We have reached desirable results using technologies of knowledge extraction, including problem statement, data preparation, and automatic construction of models, analysis and result testing, use of models in real applications. Now it is enough to address to models estimation application and after few minutes reveal the most probable «risk zones» (tax legislation infringements), react in time to possible tax offences and define necessary tax control actions.

The purpose of dissertation is the substantiation of such system construction for the tax control and developing of methodical toolkit for tax investigations to counteract tax infringements.

Achievement of an object in view has demanded the decision of following problems:

- Construction and improvement of information-analytical system of the Ministry of Finance of Georgia;

- Creation of datawarehouse. Datawarehouse allows essentially increase information systems efficiency. It can consist the information from text files and databases, both relational, and not relational, for using in decision support systems.

- Improvement of database queries performance using the materialized views, as powerful tool for increasing productivity of datawarehouses and decision support systems.

- Creation of table partitions to increase scalability of the database with huge amount of data. Using of table partitions positively affects database productivity, availability of the data and simplifies administration.
- Organization of Data Mining process, including following phases:

    1. Business understanding;
    2. Data understanding;
    3. Data preparation;

4. Modelling;
5. Evaluetion;
6. Deployment.

- Creation new methods for undisciplined taxpayers revealing and systematization of old ones;

Scientific novelty of dissertation consists in research and creation of the models, helping to find the latent dependences in big volume of the processed information. At present Using of Data Mining technology in the world market is extended enough (Large companies as ORACLE, IBM, MICROSOFT are interested in to develop Data Mining technology). However, for developing infrastructure of tax system of Georgia using given technology became a new stage of information-analytical system development, thereby providing a wide spectrum of possibilities, giving means for the situation analysis, decision-making and shows a way to meet intellectual requirements of the State Tax Service.

We have possibility to make a tax policy more purposeful and effective dividing taxpayers into various groups using Data Mining tools. Now we do not waste time on revealing unfair taxpayers from the list with thousand of rows relaying on auditors experience only.

Creation of models on which basis we can carry out the qualitative analysis of the data, occurs in the headquarters. Qualified experts can change coefficients with the subsequent data recalculation using stored procedures. After tries necessary model is found which describers the situation more precisely.Thus it is necessary to notice that unlike other methods of decision support systems, Data Mining technology possess much higher degree of intellectuality and good scalability, allowing substantially automate the data analysis.

Перечень Рисунков :

Перечень Таблиц:

Список аббревиатур

МФГ	-	Министерство Финансов Грузии,
DM	-	Data Mining,
ИС	-	Информационная система,
БД	-	База данных,
ПО	-	Программое обеспечение,
СУБД	-	Система управления базами данных,
ОС	-	Операционня система,
СППР	-	Система поддержки принятия решений,
ХД	-	Хранилище данных,
ЭВМ	-	Электронная вычислительная машина,
СОД	-	Системы обработки данных
BI	-	(англ. Business intelligence, бизнес-интеллект)
OLAP	-	(англ.Online analytical processing, аналитическая обработка в реальном времени),
SQL	-	(англ. Structured Query Language, Структуированный язык запросов)

Благодарность

Хочу выразить глубокую благодарность моему руководителю, профессору Зурабу Босикашвили, под чутким руководством которого на протяжении нескольких лет была подготовлена данная работа.

Выражаю особую благодарность г-ну Арчилу Прангишвили за оказаную помощь в обсуждении различных вопросов, связанных с диссертационной работой.

Также выражаю благодарность г-же Валиде Сесадзе и г-же Лили Иоселиани за заботу и внимание, проявленные ко мне на этапе подготовки диссертационной работы.

Благодарю м. Томаса Симсона, представителя Государственного Казначейства США, за оказанное содействие в процессе создания программого продукта.

Так же хочу поблагодарить моих коллег, принимавших участие в дискуссии и выполнении работы.

Введение

Актуальность. Налоги всегда предполагали изъятие у человека части дохода в пользу государства. Низкий уровень налоговой культуры граждан обуславливает противоречия между финансовыми потребностями государства и выполнением членами общества обязательств по полному и своевременному перечислению налогов. Многие предприниматели ищут всевозможные пути уклонения от уплаты налогов. Государство, в свою очередь, разрабатывает методы противодействия налоговым правонарушениям и преступлениям, именно поэтому актуальность проблемы разработки методов противодействия налоговым правонарушениям возрастает. Особенную актуальность приобретает формирование системы налогового контроля. Разработка методов противодействия налоговым нарушениям, методик расследования налоговых правонарушений и преступлений приобретает первостепенное значение как инструмент воздействия государства на экономическое поведение участников рыночных отношений и формирование налоговой культуры. При сложности и противоречивости существующей налоговой законодательной базы лишь высокий уровень налоговой культуры сможет обеспечить достаточную собираемость налогов. Для того чтобы эффективно проводить мероприятия налогового контроля и налогового расследования, недостаточно знать существующие способы уклонения от уплаты налогов и применять известные подходы к противодействию налоговым правонарушениям. Необходимо предвидеть возникновение новых, до сих пор неизвестных способов уклонения от уплаты налогов, для чего нужно систематизировать уже выявленные приемы недобросовестных налогоплательщиков и примененные к ним методы противодействия налоговым правонарушениям.

Исследование проблем уклонения от уплаты налогов становится актуальной общегосударственной задачей. Назрела острая необходимость консолидации усилий контролирующих органов в борьбе против теневизации экономики и налоговой асимметрии. Все вышесказанное определило выбор темы и предмета диссертационного исследования.

В настоящее время систематическое изучение проблем уклонения от уплаты налогов и расследования налоговых правонарушений в нашей стране находится на стадии становления, активно идет процесс накопления материала

по рассматриваемой проблематике, создается основа для более глубокого изучения этого явления и последующих теоретических и практических разработок.

Еще недостаточно полно отражены место и роль налогового контроля в налоговой системе и в системе финансового контроля, мало внимания уделено определению налоговой культуры, не в полной мере изучены способы уклонения от уплаты налогов и методика их расследования, а также не классифицированы методы противодействия налоговым нарушениям. Все это стало стимулом продолжения исследования актуальных аспектов проблемы.

Цель диссертационной работы. Целью диссертационной работы является теоретическое обоснование системного построения налогового контроля, разработка методического инструментария налоговых расследований и противодействия налоговым нарушениям путем исследования процессов налогообразования с помощью Data Mining методов . Необходимо отметить , что данная технология не имеет широкого применения в Грузии. Однако, ее примениние безусловно оправдывает обращение к продуктам Data Mining в различних сферах, таких как: финансы, медицина, страхование. На общем фоне развития экономики Грузии, особо остро стоит проблема рационального подхода к выявлению зон рисков налогплательщиков. Поэтому обращение налоговой службы к методам анализа посредством Data Mining, является новым, не имеющем аналогов в информационно-аналитической ситеме Министерсва Финансов, подходом к решению данной задачи.

Достижение поставленной цели потребовало решения следующих задач:

- Системно представить налоговый контроль в структуре финансового контроля государства и обосновать новый метод налогового контроля;
- Создание единой системы планирования выездных налоговых проверок;
- Создание хранилищ данных, материальных представлений с целью повышения скорости подсчетов;
- Применение Olap технологий;
- Организация процессов Data Mining;
- Разработать методику планирования выездных налоговых проверок на основе классификации способов уклонения от уплаты налогов;

- Систематизировать и предложить новые методы выявления недисциплинированных налогоплательщиков.

Предметом исследования стали отношения между экономическими субъектами в процессе взаимодействия государственных органов налогового контроля и налогоплательщиков.

Объектом исследования избраны методы выявления скрытых зависимостей элементов системы налогообложения и возникающие в ней нарушения, связанные с уклонением от уплаты налогов.

Теоретической и методологической основой исследования стали теории, гипотезы и концепции, разработанные отечественными и зарубежными учеными в области налогов, налогообложения, налогового контроля, налоговой культуры, налоговых расследований.

Методология исследования основана на использовании общенаучных методов, экономико-статистического, расчетно-конструктивного, ситуационного, экспертных оценок.

Информационной базой исследования послужили нормативно-правовые документы, регулирующие деятельность органов налогового контроля, и оперативно-розыскные мероприятия; монографическая и научная литература по налоговым преступлениям и способам борьбы с ними; публикации в периодической печати и информационно-аналитические обзоры (в том числе сети Internet); данные официальной отчетности налоговых органов; материалы научно-практических конференций на тему налогового расследования и противодействия налоговым правонарушениям и аналитические данные налоговых органов.

Положения диссертации, выносимые на защиту:

1. Повышению собираемости налогов способствует «система налогового контроля» - система объективных и субъективных критериев хозяйственной деятельности налогоплательщика, позволяющая налоговым органам наиболее эффективно проводить мероприятия налогового контроля, формирующая налоговую историю налогоплательщика. Объективными критериями являются: время существования организации, изменение численности штата, среднемесячные налоговые платежи, экономические и финансовые результаты хозяйственной деятельности. Субъективные критерии основываются на знаниях и опыте сотрудников налоговых органов, позволяющих им определить статус

каждого налогоплательщика, оперативно и эффективно избрать мероприятия налогового контроля.

Для формализации методов налоговых расследований необходима полная классификация способов уклонения от уплаты налогов: способов связанных с нелегальной хозяйственной деятельностью (сокрытие всей экономической деятельности субъекта, сокрытие отдельных экономических операций, т.е. неотражение их в бухгалтерском учете), способов связанных с сокрытием результатов хозяйственной деятельностью (коррупционные соглашения, искажение объектов налогообложения, игры обмена), способов связанных с освобождением от налоговых платежей (неправомерное использование льгот, полное освобождение от уплаты налогов, льготное освобождение от уплаты налогов, политический торг за предоставление льгот, искусственные неплатежи налогов).

Социализация общества и уровень развития налоговой культуры определяют подходы к противодействию налоговым правонарушениям. Юридический подход предполагает совершенствование налоговых норм и правил; утилитаристский (экономический) подход - акцентирование внимания на экономических интересах рационально действующих агентов; конвенциональный подход - консолидацию сил ведущих хозяйственных агентов и полномочных представителей государственной власти в борьбе с уклонением от уплаты налогов; силовой подход — ужесточение санкций против нарушителей; культурно-нормативный подход - изменение социальных норм: выработку лояльности по отношению к государству, достижение уважения к формальным правилам, преодоление стереотипных представлений о невозможности их соблюдения и воспитание налоговой культуры налогоплательщиков.

Научная новизна диссертационного исследования:

- раскрыта сущность системы налогового контроля как нового метода налогового контроля, формирующего налоговую историю налогоплательщиков по объективным и субъективным критериям оценки их хозяйственной деятельности;
- предложено применение методов Data Mining, при построении моделей для выявления недобросовестных налогоплательщиков. Модели строятся автоматически на основе анализа имеющихся данных об объектах,

наблюдениях и ситуациях с помощью специальных алгоритмов. Основу опции Data Mining составляют процедуры, реализующие различные алгоритмы построения моделей классификации и деревьев решений. На этапе подготовки данных обеспечивается доступ к любым реляционным базам, текстовым файлам.

- Дополнительные средства преобразования и очистки данных позволяют изменять вид представления, проводить нормализацию значений, выявлять неопределенные или отсутствующие значения. На основе подготовленных данных специальные процедуры автоматически строят модели для дальнейшего прогнозирования, классификации новых ситуаций, выявления аналогий. Графические средства предоставляют широкие возможности для анализа полученных результатов, верификации моделей на тестовых наборах данных, оценки точности и устойчивости результатов. Важной особенностью созданного нами продуктаявляются его технические характеристики: работа в архитектуре клиент-сервер, использование техники параллельных вычислений, высокая степень масштабируемости при увеличении вычислительных ресурсов.

Практическую значимость диссертационного исследования имеет предложенная методика расследования налоговых правонарушений и преступлений на основе авторской классификации способов уклонения от уплаты налогов, а также разработанные схемы каждого из рассмотренных способов уклонения от уплаты налогов. Данная методика может быть использована в работе налоговых органов.

Диссертационная работа “ Разработка и исследование моделей оценок системы налогового контроля ” состоит из введения, четырех глав, заключения. Общий объем работы составляет 114 страниц: основное содержание работы отражено на 100 страницах печатного текста, 13 рисунков, библиография охватывает 91 источник.

Глава 1
Применение аналитических технологий в системе разработки налогового контроля.

1.1 Обзор

На рынке программного обеспечения Data Mining существует огромное разнообразие продуктов, относящихся к этой категории. Для выбора продукта следует тщательно изучить поставленные задачи и обозначить те результаты, которые необходимо получить. Существуют различные варианты решений по внедрению инструментов Data Mining, например:

- покупка готового программного обеспечения Data Mining;
- покупка программного обеспечения Data Mining, адаптированного под конкретный бизнес;
- разработка Data Mining-продукта на заказ сторонней компанией;
- разработка Data Mining-продукта своими силами;
- различные комбинации вариантов, описанных выше, в том числе использование различных библиотек, компонентов и инструментальные наборы для разработчиков создания встроенных приложений Data Mining.

В начале 90-х годов прошлого столетия рынок Data Mining насчитывал около десяти поставщиков. В средине 90-х число поставщиков, представленных компаниями малого, среднего и большого размера, насчитывало более 50 фирм. Сейчас к аналитическим технологиям, в том числе к Data Mining, проявляется огромный интерес. На этом рынке работает множество фирм, ориентированных на создание инструментов Data Mining, а также комплексного внедрения Data Mining, OLAP и хранилищ данных. Инструменты Data Mining во многих случаях рассматриваются как составная часть BI-платформ, в состав которых также входят средства построения хранилищ и витрин данных, средства обработки неожиданных запросов (ad-hoc query), средства отчетности (reporting), а также инструменты OLAP.

Разработкой в секторе Data Mining всемирного рынка программного обеспечения заняты как всемирно известные лидеры, так и новые развивающиеся компании. Инструменты Data Mining могут быть представлены

либо как самостоятельное приложение, либо как дополнения к основному продукту. Последний вариант реализуется многими лидерами рынка программного обеспечения. Так, уже стало традицией, что разработчики универсальных статистических пакетов, в дополнение к традиционным методам статистического анализа, включают в пакет определенный набор методов Data Mining. Это такие пакеты как SPSS (SPSS, Clementine), Statistica (StatSoft), SAS Institute (SAS Enterprise Miner). Некоторые разработчики OLAP-решений также предлагают набор методов Data Mining, например, семейство продуктов Cognos. Есть поставщики, включающие Data Mining решения в функциональность СУБД: это Microsoft(Microsoft SQL Server), Oracle, IBM .Интересными являются данные опроса "Инструменты Data Mining, которые Вы регулярно используете", проведенного в мае 2005 года на Kdnuggets. Результаты представлены на рисунке [75]

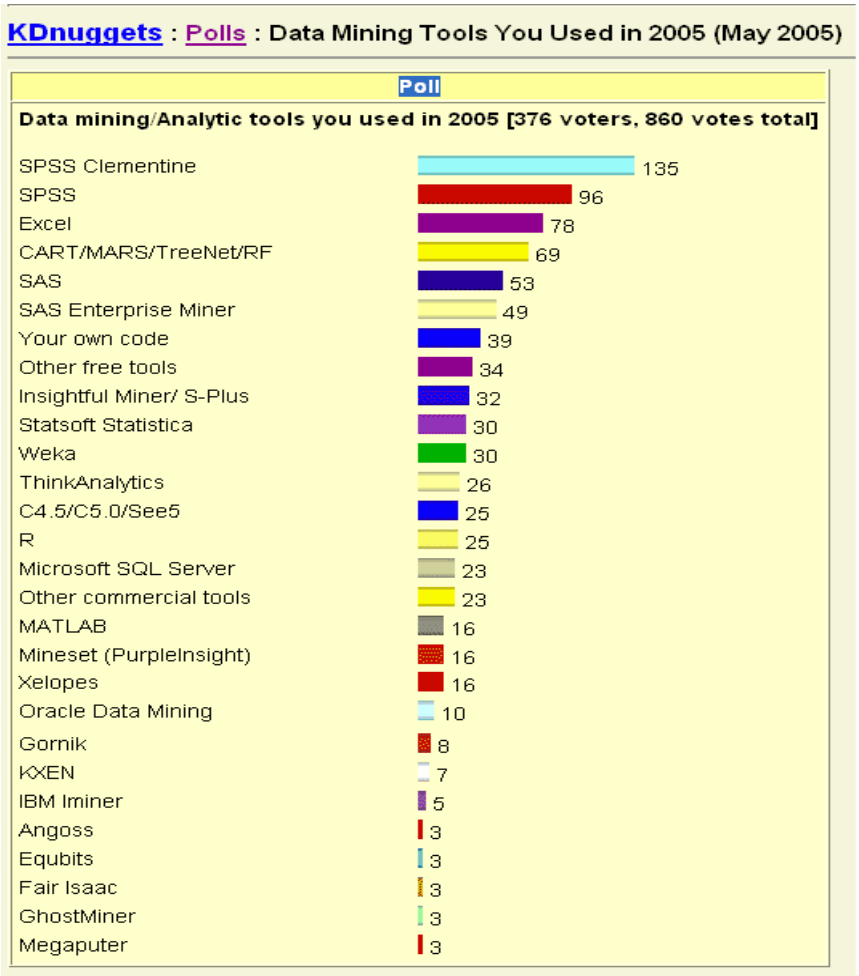

Рис.1 Инструменты Data Mining

Различные варианты внедрения Data Mining имеют свои сильные и слабые стороны. Так, преимуществами готового программного обеспечения являются

готовые алгоритмы, техническая поддержка производителя, полная конфиденциальность информации, а также не требуется дописывать программный код, существует возможность приобретения различных модулей и надстроек к используемому пакету, общение с другими пользователями пакета и др.

Однако, такое решение имеет и слабые стороны. В зависимости от инструмента, это может быть достаточно высокая стоимость лицензий на программное обеспечение, невозможность добавлять свои функции, сложность подготовки данных, практическое отсутствие в интерфейсе терминов предметной области и другие. Такое решение требует наличия высококвалифицированных кадров, которые смогут качественно подготовить данные к анализу, знают, какие алгоритмы следует применять для решения каких задач, сумеют проинтерпретировать полученные результаты в терминах решаемых бизнес-задач. Далеко не каждая компания может содержать штат таких специалистов, а зачастую их содержание даже неэффективно. Для эффективного применения технологии Data Mining требуются квалифицированные специалисты, которые сумеют качественно провести весь цикл анализа. Пока что таких грамотных специалистов не только в нашем государстве, но и на просторах СНГ очень немного, и потому они довольно дороги. Готовые продукты изначально предназначены для решения хотя и широкого, но все же стандартного и ограниченного круга задач - адаптация продукта к условиям конкретного бизнеса ложится на плечи сотрудников компании.

Большая стоимость , дорогостояшие услуги высококвалифицированных специалистов, а так же немаловажний фактор - большие сроки внедрения, побудили нас решится на создание аналитической системы выявления зон рисков налогплательщиков самостоятельно. При тесном сотрудничестве и поддержке представителей Государственного Казначейства Соединенных Штатов Америки Мы сформулировали цели , поставленной перед нами задачи, обсудили наиболее вероятно возникающие проблемы, в соответствии с последовательностью работ, выполняемых в рамках аналитического процесса.

Одной из главних проблем была разрозненность данных. Раньше каждая региональная налоговоя инспекция имела свой собственный сервер, что представляло собой преграду при обмене информацией между отдельными

управлениями системы. управление базой данных стало гораздо проще при переходе на единый сервер, таким образом аднимистрирование данными проиходит в налоговом департаменте а не в каждой инспекции в отдельности.

Далее мы столкнулись с проблемой, когда интеграции подлежит множество источников данных, содержаших разрозненные данные в различном представлении. Раньше часть банковских платежей, а так же информацию по таможенным декларациям прироходилось загружать путем импорта из файлов Excel и Dbf, данные налогового аудита импортировались из базы Access. Таким образом, перед нами встала необходимость консолидации различных представлений данных и исключение дублирующейся информации для обеспечения доступа к точным и согласованным данным.

Так же немаловажной проблемой является качество данных: Для обеспечения качественного анализа необходимо проведение предварительной обработки данных, которая является необходимым этапом процесса Data Mining. Данные высокого качества - это полные, точные, своевременные данные, которые поддаются интерпретации.

Небходимо отметить, важность использования технологий хранилищ данных как информационной основы для Data Mining. Структура хранилища, оптимизированная под задачи аналитической обработки, позволяет свести к минимуму потери времени на поиск нужных данных и получение промежуточных результатов. Наличие в организации хранилища данных делает анализ проще и эффективней. Необходимо отметить , что эффективность Хранилища Данных напрямую зависит от его производительности. На производительность Хранилища способны влиять операции баз данных. Если другие базы данных находятся на том же сервере, что и Хранилище, то их работа потребует определенных ресурсов сервера. Для того чтобы обеспечить оптимальную производительность как приложений по обработке транзакций, так и Хранилища, они должны размещаться на разных серверах. Если необходимо, чтобы базы данных транзакций и Хранилища находились на одном сервере, то это должны быть разные экземпляры баз данных для того, чтобы можно было использовать СУБД и параметры настройки баз данных, относящиеся к каждой из них отдельно. В зависимости от изменений, связанных с ростом и использованием Хранилища, экземпляры СУБД и баз данных могут потребовать дополнительной настройки для поддержки их производительности.

Так же для оприимизации запросов в Хранилищах Данных мы применяем методы индексирования и фрагментации таблиц.

Для повышения производительности в хранилищах данных и системах поддержки принятия решений используют материализованные представления, которые многократно ускоряют выполнение запросов, обращающихся к большому количеству записей. Это достигается за счет прозрачного использования заранее вычисленных итоговых данных и результатов соединений таблиц.[22]

Существует несколько профилактических операций наблюдения, которые необходимо осуществлять в среде работы Хранилища данных для того, чтобы избежать общих нарушений, которые могут влиять на производительность и доступность Хранилища. Области, в которых следует проводить подобные проверки, включают системы файлов, базы данных, сектора временного хранения данных, табличные пространства баз данных, журнальные файлы и архивы. Также необходимо оценивать загруженность центрального процессора и памяти, время ожидания ввода-вывода, скорость передачи данных в каналах сетевой связи, скорость буферной памяти (т.е. скорость, с которой данные передаются из памяти на диск) и т.д. В зависимости от размера, избыточности, способности восстанавливаться после отказа, общего распределения и гибкости среды Хранилища данных такой мониторинг может потребоваться на многих уровнях и серверах инфраструктуры. [27]

Успешный анализ требует очищенных и подготовленных данных. По утверждению аналитиков и пользователей очистка клиентских записей, разделение их на поддающиеся обработке, но статистически достоверные образцы, а затем тестирование и уточнение всех результатов, занимает до 80% процентов всего DM-процесса. Таким образом, чтобы заставить технологию работать на себя потребуется много времени. Много усилий тратится на анализ предварительных данных и корректировку прогнозирующих моделей. Так же немаловажной проблемой является качество данных. Для обеспечения качественного анализа необходимо проведение предварительной обработки данных, которая является необходимым этапом процесса Data Mining. Данные высокого качества - это полные, точные, своевременные данные, которые поддаются интерпретации.

Очистка данных занимается выявлением и удалением ошибок и несоответствий в данных с целью улучшения качества данных.

Проблемы с качеством встречаются в отдельных наборах данных - таких как файлы и базы данных. Когда интеграции подлежит множество источников данных (например в Хранилищах, интегрированных системах баз данных или глобальных информационных Интернет-системах), необходимость в *очистке данных* существенно возрастает. Это происходит оттого, что источники часто содержат разрозненные данные в различном представлении. Для обеспечения доступа к точным и согласованным данным необходима консолидация различных представлений данных и исключение дублирующейся информации. Специальные средства очистки обычно имеют дело с конкретными областями - в основном это имена и адреса - или же с исключением дубликатов. Преобразования обеспечиваются либо в форме библиотеки правил, либо пользователем в интерактивном режиме.

Метод очистки данных должен удовлетворять ряду критериев .

1. Он должен выявлять и удалять все основные ошибки и несоответствия, как в отдельных источниках данных, так и при интеграции нескольких источников.
2. Метод должен поддерживаться определенными инструментами, чтобы сократить объемы ручной проверки и программирования, и быть гибким в плане работы с дополнительными источниками.
3. Очистка данных не должна производиться в отрыве от связанных со схемой преобразования данных, выполняемых на основе сложных метаданных.
4. Функции маппирования для очистки и других преобразований данных должны быть определены декларативным образом и подходить для использования в других источниках данных и в обработке запросов.
5. Инфраструктура технологического процесса должна особенно интенсивно поддерживаться для Хранилищ данных, обеспечивая эффективное и надежное выполнение всех этапов преобразования для множества источников и больших наборов данных.

На сегодняшний день интерес к очистке данных возрастает. Целый ряд исследовательских групп занимается общими проблемами, связанными с очисткой данных, в том числе, со специфическими подходами к Data Mining и

преобразованию данных на основании сопоставления схемы. В последнее время некоторые исследования коснулись единого, более сложного подхода к очистке данных, включающего ряд аспектов преобразования данных, специфических операторов и их реализации.

В целом, очистка данных включает следующие этапы :

1. Анализ данных.
2. Определение порядка и правил преобразования данных.
3. Подтверждение.
4. Преобразования.
5. Противоток очищенных данных.

Этап № 1. Анализ данных.

Подробный анализ данных необходим для выявления подлежащих удалению видов ошибок и несоответствий. Здесь можно использовать как ручную проверку данных или их шаблонов, так и специальные программы для получения метаданных о свойствах данных и определения проблем качества.

Этап № 2. Определение порядка и правил преобразования данных.

В зависимости от числа источников данных, степени их неоднородности и загрязненности, данные могут требовать достаточно обширного преобразования и *очистки*. Иногда для отображения источников общей *модели данных* используется трансляция схемы; для Хранилищ данных обычно используется реляционное представление. Первые шаги по *очистке* могут уточнить или изменить описание проблем отдельных источников данных, а также подготовить данные для интеграции. Дальнейшие шаги должны быть направлены на интеграцию схемы/данных и устранение проблем множественных элементов, например, дубликатов. Для Хранилищ в процессе работы должны быть определены методы контроля и поток данных, подлежащий преобразованию и очистке.

Преобразования данных, связанные со схемой, так же как и этапы очистки, должны, насколько возможно, определяться с помощью декларативного запроса и языка маппирования, обеспечивая, таким образом, автоматическую генерацию кода преобразования. К тому же, в процессе преобразования должна существовать возможность запуска написанного пользователем кода очистки и специальных средств. Этапы преобразования

могут требовать обратной связи с пользователем по тем элементам данных, для которых отсутствует встроенная логика очистки.

Этап № 3. Подтверждение.

На этом этапе определяется правильность и эффективность процесса и определений преобразования. Это осуществляется путем тестирования и оценивания, например, на примере или на копии данных источника, - чтобы выяснить, необходимо ли как-то улучшить эти определения. При анализе, проектировании и подтверждении может потребоваться множество итераций, например, в связи с тем, что некоторые ошибки становятся заметны только после проведения определенных преобразований.

Этап № 4. Преобразования.

На этом этапе осуществляется выполнение преобразований либо в процессе обновления Хранилища данных, либо при ответе на запросы по множеству источников.

Этап № 5. Противоток очищенных данных.

После того как ошибки отдельного источника удалены, загрязненные данные в исходных источниках должны замениться на очищенные, для того чтобы улучшенные данные попали также в унаследованные приложения и в дальнейшем при извлечении не требовали дополнительной очистки. Для Хранилищ очищенные данные находятся в области хранения данных.

Такой процесс преобразования требует больших объемов метаданных (схем, характеристик данных уровня схемы, определений технологического процесса и др.). Для согласованности, гибкости и упрощения использования в других случаях, эти метаданные должны храниться в репозитории на основе СУБД. Для поддержки *качества данных* подробная информация о процессе преобразования должна записываться как в репозиторий, так и в трансформированные элементы данных, в особенности информация о полноте и свежести исходных данных и происхождения информации о первоисточнике трансформированных объектов и произведенных с ними изменениях.

Мы избавились от распространенных видов загрязнения следующим образом: очистили их от дубликатов, путем замены группы дубликатов на одну уникальную запись, заменили пропушенные значения, путем замены пропущенных значений на возможные значения.

Наш продукт поддерживает все этапы технологии извлечения знаний, включая постановку задачи, подготовку данных, автоматическое построение моделей, анализ и тестирование результатов, использование моделей в реальных приложениях.

Модели строятся автоматически на основе анализа имеющихся данных об объектах, наблюдениях и ситуациях с помощью специальных алгоритмов. Основу опции Data Mining составляют процедуры, реализующие различные алгоритмы построения моделей классификации и деревьев решений. На этапе подготовки данных обеспечивается доступ к любым реляционным базам, текстовым файлам.

Дополнительные средства преобразования и очистки данных позволяют изменять вид представления, проводить нормализацию значений, выявлять неопределенные или отсутствующие значения. На основе подготовленных данных специальные процедуры автоматически строят модели для дальнейшего прогнозирования, классификации новых ситуаций, выявления аналогий. Графические средства предоставляют широкие возможности для анализа полученных результатов, верификации моделей на тестовых наборах данных, оценки точности и устойчивости результатов.

Важной особенностью созданного нами продуктаявляются его технические характеристики: работа в архитектуре клиент-сервер, использование техники параллельных вычислений, высокая степень масштабируемости при увеличении вычислительных ресурсов.

Необходимо отметить, что опцию в Data Minig входят средства подготовки данных, оценки результатов, применения моделей к новым наборам данных. Использовать все эти возможности можно как на программном уровне с помощью PL/SQL API, так и с помощью графической среды, которая ориентирована на работу аналитиков, решающих задачи прогнозирования, выявления тенденций, сегментации и другие.

Обычно в программу загружаются данные, необходимые для дальнейшего анализа. После получения выборки можно получить подробную статистику по ней, посмотреть, как выглядят данные на диаграммах.

После такого разведочного анализа можно принимать решения о необходимости предобработки данных. Например, если статистика показывает,

что в выборке есть пустые значения (пропуски данных), можно применить фильтрацию для их устранения.

Предобработанные данные далее подвергаются трансформации. Например, нечисловые данные преобразуются в числовые, что необходимо для некоторых алгоритмов, а так же удаляются дубликаты.

К трансформированным данным применяются методы более глубокого анализа. На этом этапе выявляются скрытые зависимости и закономерности в данных, на основании которых строятся различные модели. Модель представляет собой шаблон, который содержит формализованные знания.

Последний этап - интерпретация - предназначен, чтобы из формализованных знаний получить знания на языке предметной области.

Вся работа по анализу данных в нашем продукте базируется на выполнении следующих действий:импорт данных;

- обработка данных;
- визуализация;
- экспорт данных.

Итогом работ по интеллектуальному анализу данных является развертывание созданной модели - это заключительная стадия, на которой реализуется экономическая отдача от проведенных исследований.

Основные характеристики созданного нами продукта :

Простой графический интерфейс, создающий диаграммы процессов обработки данных:

- Быстрое создание большого числа качественных моделей.
- Возможность доступа через Web-интерфейс.

Масштабируемая обработка

- Серверная обработка
- Параллельная обработка
- Все хранение и обработка данных - на серверах.

В заключение хочется сказать, что область использования Data Mining ничем не ограничена - она везде, где имеются какие-либо данные. Отметим, что на сегодняшний день наибольшее распространение технология Data Mining получила при решении бизнес-задач. Сейчас технология Data Mining используется практически во всех сферах деятельности человека, где накоплены

ретроспективные данные, это наука, бизнес, исследования для правительства и Web-направление.

Выбор инструментального средства Data Mining и способа его внедрения должен проводиться в соответствии с конкретными целями и задачами, учитывать уровень финансовых возможностей компании, квалификацию пользователей и целый ряд других факторов. Поскольку внедрение Data Mining почти всегда требует серьезных финансовых затрат. Также следует не только учитывать задачи, которые стоят перед компанией сегодня, но и рассчитывать на возможность возникновения новых задач в перспективе.

1.2 Налоговый контроль в налоговой системе, его цели и задачи.

Эффективное функционирование любого государства и стабильное финансирование предусмотренных бюджетами мероприятий требует систематического пополнения финансовых ресурсов на федеральном, региональном и местном уровнях. Это достигается в основном за счет уплаты юридическими и физическими лицами налогов и других обязательных платежей. В соответствии с действующим налоговым законодательством и другими нормативными актами плательщики обязаны уплачивать указанные платежи в установленных размерах и в определенные сроки. В широком смысле под налогами понимаются обязательные платежи в бюджет, осуществляемые юридическими и физическими лицами.

Под налогом понимается обязательный, индивидуально безвозмездный платеж, взимаемый с организаций и физических лиц в форме отчуждения принадлежащих им на праве собственности, хозяйственного ведения или оперативного управления денежных средств, в целях финансового обеспечения деятельности государства и/или муниципальных образований.

Налоги - это один из экономических рычагов, при помощи которых государство воздействует на рыночную экономику. В условиях рыночной экономики любое государство широко использует налоговую политику в качестве определенного регулятора воздействия на негативные явления рынка. Налоги, как и вся налоговая система, являются мощным инструментом управления экономикой в условиях рынка.Но, к сожалению, на практике юридические и физические лица допускают несвоевременную уплату налогов и других обязательных платежей в связи с рядом объективных и субъективных причин. С переходом к рыночным отношениям создаются новые предприятия, осуществляющие свою финансово-хозяйственную деятельность в различных сферах экономики. Многие из них не имеют достаточно квалифицированных специалистов в области бухгалтерского учета. На таких предприятиях, как правило, допускаются ошибки в учете. Нередки случаи сознательного искажения отчетных данных. Причем сегодня стало естественным уклонение от налоговой повинности, как легальными когда удается полностью или частично избежать налогообложения, не нарушая при этом действующего

законодательства, - так и нелегальными, то есть запрещенными законом способами.

Все это приводит к занижению налогооблагаемой базы и недопоступлению в бюджет налогов и других приравненных к ним платежей. Ошибки в исчислении и уплате налогов допускаются также из-за частых изменений в законодательстве.

В связи с этим сегодня перед налоговыми органами встает серьезная проблема - контроль за правильностью, своевременностью и полнотой взимания налогов и его совершенствование.

1.3 Автоматизированная информационная система и ее компоненты.

Практически общепринятой в настоящее время стала концепция построения информационных систем на основе реляционной модели данных. В пользу выбора этой модели для построения и рассматриваемой нами говорят следующие обстоятельства:

- реляционная модель хорошо исследована, для нее выработаны приемы и методы использования, позволяющие решать практически любые задачи хранения данных и доступа к ним, разработаны также методы приведения к реляционной модели тех данных, предметная структуризация которых естественным образом в реляционную модель не вписывается;
- реляционная модель интуитивно понятна как разработчику, так и конечному пользователю, так как ее прообразом являются таблицы - хорошо знакомый всем инструмент;
- практически все промышленно выпускаемые на сегодняшний день средства управления базами данных поддерживают реляционную модель;
- для реляционной модели существует мощное средство формулирования запросов к базе данных - структурированный язык запросов SQL. Являясь языком непроцедурным, SQL, таким образом, не зависти от среды (аппаратной и операционной), в которой он выполняется. SQL является де-факто стандартом обращений к базам данных, стандарт ANSI SQL поддерживается ISO и обеспечивается большинством промышленно выпускаемых средств.

Из наиболее популярных современных многопользовательских СУБД следует назвать:

- MS SQL Server фирмы Microsoft;
- Oracle фирмы Oracle;
- DB2 фирмы IBM.

Функциональные возможности названных СУБД практически одинаковы: все они обеспечивают язык SQL, как средство формулирования запросов, обеспечивают весь необходимый сервис для администрирования базы

данных, работу СУБД в режиме клиент/сервер с параллельным многопользовательским доступом к данным. Следует, однако, заметить, что MS SQL Server проигрывает по сравнению с двумя другими названными СУБД в отношении эффективности выполнения при равных ресурсах. Еще одним недостатком этой СУБД следует считать то, что она работает только в среде операционной системы Windows NT, что ограничивает ее применимость только персональной платформой (пусть даже и мощными персональными серверами). Что касается DB2 и Oracle, то эти СУБД принадлежат к числу первых реляционных СУБД и, следовательно, имеют наиболее богатую историю развития и совершенствования. Именно DB2 явилась результатом того проекта корпорации IBM, в котором была сформулирована реляционная модель данных и разработан язык SQL. Обе эти СУБД являются многоплатформенными, хотя Oracle ориентирован, прежде всего, на выполнение в среде операционной системы UNIX и в UNIX-подобных системах. [75]

DB2, однако, адаптирована к большему разнообразию аппаратных и операционных сред. По эффективности две указанные СУБД конкурируют между собой с переменным успехом, но Oracle предъявляет большие требования к ресурсам при равной производительности, поэтому стоимость транзакции в DB2 получается на 15-20% ниже.

Уже давно наступило время, когда под автоматизацией предприятий стало подразумеваться не просто приобретение компьютеров и создание корпоративной сети, но создание информационной системы, включающей в себя и компьютеры, и сети, и программное обеспечение, а главное - организацию информационных потоков. Проанализировав опыт внедрения информационных систем на отечественных предприятиях, можно заметить, что время от времени ИС на базе какого-либо интегрированного продукта либо внедряются не до конца, либо руководство компаний ими практически не пользуется.

Анализ внедрений, осуществленных на сегодняшний день, выявляет несколько причин неудач при создании ИС:

1. Первая состоит в том, что готовые западные системы ориентированы на некие идеальные бизнес-процессы, оторванные от реальной структуры конкретной компании. А реальные учреждения, компании и корпорации вовсе не идеальны, а наоборот, очень сложны с точки зрения иерархии управления.

Более того, зачастую формальная иерархия причудливо переплетается с реальной.

2. Вторая причина - в том, что исторически разработкой систем занимались программисты, в силу чего они строились согласно теории автоматизированных систем. Получался замкнутый автоматизированный процесс, по возможности исключающий человека. В результате весь средний менеджмент такой системой отторгался. Поэтому руководители среднего звена противятся внедрению таких систем и сознательно, и бессознательно.

3. Третье - это недостаточный анализ существующих задач на этапе проектирования. Например, на Западе, в частности, в США, у компаний-заказчиков, как правило, есть специальные отделы, которые планируют работы по автоматизации и анализируют: что надо автоматизировать, что не надо, что выгодно, а что убыточно, и как вообще должна быть построена система, какие функции она должна выполнять. У отечественных компаний подобные структуры, как правило, отсутствуют.

Сегодня необходим новый подход к созданию информационных систем. Новизна заключается не в создании системы на базе какого-либо интегрированного продукта, а в тщательном проектировании системы и лишь потом реализации ее с помощью адекватных программных средств.

Не секрет, что зачастую подход к автоматизации бывает таким: нужно автоматизировать все, а поэтому покупаем могучую интегрированную систему и модуль за модулем всю ее внедряем. Но уже потом выясняется, что полученный эффект весьма далек от ожидаемого и деньги потрачены впустую. На практике для решения конкретной проблемы компании бывает достаточно иметь электронную почту и Excel. Иногда бывает нужно внедрить всего лишь несколько специализированных и недорогих приложений и связать их на базе интеграционной платформы или там, где это необходимо, использовать функциональность ERP-системы. Все эти вопросы можно и нужно решать на этапе проектирования, т. е. осознанно подходить к выбору средств автоматизации, сравнивая затраты с ожидаемым эффектом.

Нынешних огрехов проектирования можно избежать, используя принцип, который называется *синархическим проектированием*. Этот новый принцип является проявлением "закона синархии", который описал в начале XX века

российский философ Владимир Шмаков. Если кратко, то это органичное сочетание определенной иерархии и аналогии в построении мироздания.[54]

Синархическое проектирование - это технология, которая позволяет создавать ИС для конкретного предприятия, холдинга или концерна с учетом реальной иерархии управления, поэтапно ее внедрять, реально планировать и получать эффект от внедрения на каждом этапе, органично встраивать в систему стандартные компоненты и оригинальные разработки. Более того, синархическое проектирование позволяет овладеть системой как инструментом управления на всех уровнях - от исполнителя до директора. При этом ответственность не перекладывается на систему, и руководителю понятно происхождение информации, в ней циркулирующей.

В заключение необходимо подчеркнуть, что и заказчику, и поставщику решения еще до выбора того или иного ПО для создания ИС необходимо, прежде всего, провести анализ, что им действительно необходимо автоматизировать, после чего заняться проектированием. Другими словами, только тщательное предпроектное обследование, а затем проектирование с учетом всех особенностей реальной структуры управления конкретной компании дадут в итоге действительный эффект от внедрения автоматизированной информационной системы, к которому в конечном итоге стремятся и заказчики, и системные интеграторы.

Таким образом, автоматизированная информационная система предназначена для полномасштабной автоматизации управления информационными ресурсами , их интеграции в единое информационное пространство, обеспечения полного доступа к ним со стороны пользователей, и обеспечения информационного обмена между различными региональными налоговыми инспекциями и министрерством финансов Грузии.

В АИС за хранение информации отвечают:

- на физическом уровне
    - встроенные устройства памяти
    - внешние накопители
    - дисковые массивы
- на программном уровне
    - файловая система ОС
    - СУБД

- Системы хранения документов, мультимедиа и т. д.

На рисунке схематически представлена структура Базы Данных Министерства Финансов Грузии.

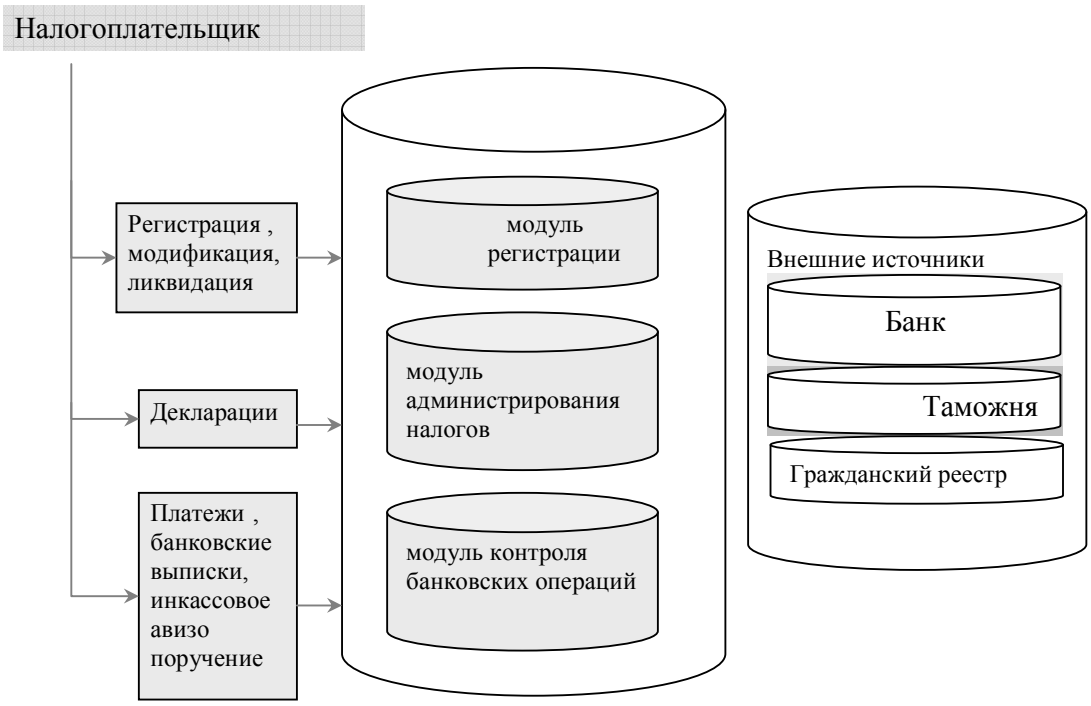

Рис 2. Информационные источники МФГ

- о зарегистрированних юридических и физических лицах;
- о банковских счетах организаций и физических лиц;

модуль аднинистрирования налогов, куда поступает информация;

- предъявленных налогоплательщиками декларациях;
- информация о банковских блатежах;

информация из таможенной базы данных;

А так же гражданский реестр, содержаший информацию о паспортных данных физических лиц, необходимую при регистрации.

Т.е. мы владеем информацией о регистрации налогоплательшика, банковских платежах, декларациях, информацией о таможенном и налоговом контроле.

Разобраться в таком большом, постоянно пополняюшемся обьеме, сырых данних достаточно сложно. Для принятия решения о выборке

налогоплательщиков, которые требуют повышенного внимания , появляется необходимость систематизации данных. Стремление к повышению эффективности использования информации побудило нас обратиться к технологиям Data Mining, тема которых подробнее раскрывается в следующих главах.

1.4 Исследование возможоностей Data Mining с целью выявления зон риска налогоплательщиков

Высокая значимость налоговых поступлений в формировании финансовых ресурсов государства определяет особую роль государственного налогового контроля. Вместе с тем задолженность по налоговым платежам юридических и физических лиц составила на начало 2006 года около 30% общего объема налоговых поступлений в консолидированный бюджет. Свыше 44% налогоплательщиков, состоящих на учете в налоговых органах, не предоставили налоговую отчетность за 2007 год. Неопределенности и несоответствия в налоговом законодательстве способствуют созданию многочисленных схем ухода от уплаты налогов, наиболее массовыми из них являются возмещение НДС, «зависшие платежи» в «проблемных» банках, преднамеренное банкротство организаций и др. В то же время эффективность выездных налоговых проверок налогоплательщиков невысока: практически каждая вторая из них заканчивается безрезультатно, что свидетельствует о необходимости совершенствования методологии, методики и организации налогового контроля с учетом особенностей формирования налоговой базы . Прежде всего требует разработки методология отбора налогоплательщиков с целью проведения камеральных проверок. Качественного совершенствования требует методика выездных налоговых проверок, начиная с выбора объекта проверки до оформления и реализации её результатов. Для повышения результативности контрольной работы налоговых органов требуется разработка эффективных методов ведения налоговых проверок с целью выявления - типичных схем ухода от налогообложения и сокрытия налоговой базы с учетом организационных и отраслевых особенностей налогоплательщиков. Отсутствие законодательно утвержденного регламента обмена информацией налоговых органов с органами налоговой полиции снижает результаты налогового контроля за правильностью исчисления, полнотой и своевременностью уплаты .

Проблема совершенствования налогового контроля охватывает множество аспектов, и в первую очередь, с точки зрения информационных технологий необходимо - разработать и научно обосновать предложения по совершенствованию организации и методов контрольной работы налоговых органов и определить критерии оценки ее эффективности можно с помощью

использования новых технологий , например, таких как хранилиша данних применение технилогий OLAP и Data Mining.

Старые методы, применявшиеся математиками и статистиками, отнимали много времени, чтобы в результате получить конструктивную и полезную информацию.

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей. Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялся в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

Data Mining - мультидисциплинарная область, возникшая и развивающаяся, в основном, на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных.

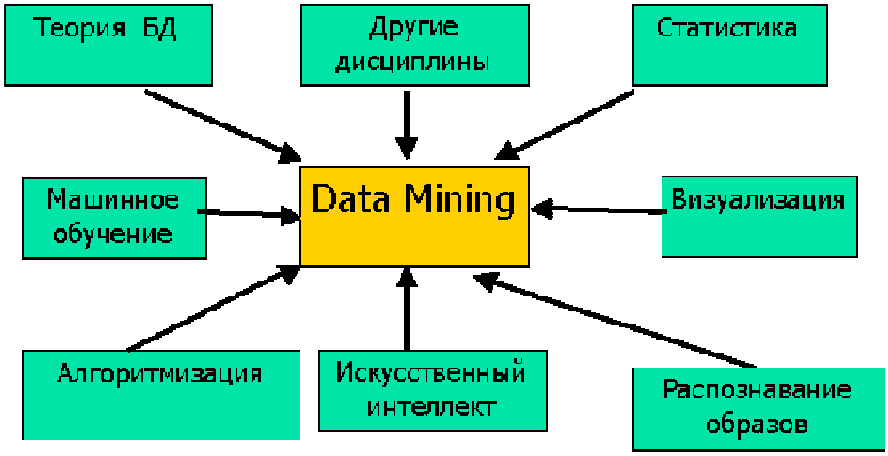

Рис. 3 Data Mining как мультидисциплинарная область

Приведем краткое описание следующих дисциплин, на стыке которых появилась новая технология, дает общее представление о продукте Data Mining.

Статистика - это наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению.

Статистика является совокупностью методов планирования эксперимента, сбора данных, их представления и обобщения, а также анализа и получения выводов на основании этих данных.

Статистика оперирует данными, полученными в результате наблюдений либо экспериментов.

Что же касается единого определения машинного обучения, то на сегодняшний день его нет.

Машинное обучение можно охарактеризовать как процесс получения программой новых знаний. Митчелл в 1996 году дал такое определение: "Машинное обучение - это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы".

Одним из наиболее популярных примеров алгоритма машинного обучения являются нейронные сети.

Искусственный интеллект - научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными. Термин интеллект (intelligence) происходит от латинского intellectus, что означает ум, рассудок, разум, мыслительные способности человека.

Соответственно, искусственный интеллект толкуется как свойство автоматических систем брать на себя отдельные функции интеллекта человека. Искусственным интеллектом называют свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека.

Каждое из направлений, сформировавших Data Mining, имеет свои особенности.

- Статистика
    - Более, чем Data Mining, базируется на теории.
    - Более сосредоточивается на проверке гипотез.
- Машинное обучение
    - Более эвристично.
    - Концентрируется на улучшении работы агентов обучения.
- Data Mining.
    - Интеграция теории и эвристик.

- Сконцентрирована на едином процессе анализа данных, включает очистку данных, обучение, интеграцию и визуализацию результатов.

Данное сравнение показывает , что Data Mining охватывает возможности составляющих компонентов выше перечисленных дисциплин.

Понятие Data Mining тесно связано с технологией баз данных и понятием данные. В течение периода эволюции баз данных многие исследователи экспериментировали с новым подходом в направлениях структуризации баз данных и обеспечения к ним доступа. Целью этих поисков было получение реляционных прототипов для более простого моделирования данных. В результате, в 1985 году был создан язык, названный SQL. На сегодняшний день практически все СУБД обеспечивают данный интерфейс.

Возникновение и развитие Data Mining обусловлено различными факторами, основные среди которых являются следующие:

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи *данных*;
- накопление большого количества ретроспективных *данных*;
- совершенствование алгоритмов обработки информации.

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных - это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

Знания - совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете или проблеме.

Полученные знания применяются для улучшения принятия бизнес решений путем использования систем поддержки принятия решений. Business

Intelligence(бизнес-интеллект) - программные средства, функционирующие в рамках предприятия и обеспечивающие функции доступа и анализа информации, которая находится в хранилище данных, а также обеспечивающие принятие правильных и обоснованных управленческих решений.

Понятие BI объединяет в себе различные средства и технологии анализа и обработки данных масштаба предприятия.

На основе этих средств создаются BI-системы, цель которых - повысить качество информации для принятия управленческих решений.

BI-системы также известны под названием Систем Поддержки Принятия Решений (СППР, DSS, Decision Support System). Эти системы превращают данные в информацию, на основе которой можно принимать решения, т.е. поддерживающую принятие решений.

Gartner Group определяет состав рынка систем Business Intelligence как набор программных продуктов следующих классов:

- средства построения хранилищ данных (data warehousing, ХД);
- системы оперативной аналитической обработки (OLAP);
- информационно-аналитические системы (Enterprise Information Systems, EIS);
- средства интеллектуального анализа данных (data mining);
- инструменты для выполнения запросов и построения отчетов (query and reporting tools).

Извлечение полезных сведений невозможно без хорошего понимания сути данных, а успешный анализ требует качественной предобработки данных. По утверждению аналитиков и пользователей баз данных, процесс предобработки может занять до 80% процентов всего Data Mining-процесса.[77]

Таким образом, чтобы технология работала на себя, необходимо приложить много усилий и времени, которые уходят на предварительный анализ данных, выбор модели и ее корректировку.

С помощью Data Mining можно отыскивать действительно очень ценную информацию, которая вскоре даст большие дивиденды в виде финансовой и конкурентной выгоды.

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии.

Области, где применения технологии Data Mining, скорее всего, будут успешными, имеют такие особенности:

- требуют решений, основанных на знаниях;
- имеют изменяющуюся окружающую среду;
- имеют доступные, достаточные и значимые данные;
- обеспечивают высокие дивиденды от правильных решений.

Поэтому наилучшее применение Data Mining возможно в сфере работы налогового контроля для выявления лиц уклоняющихся от налогов. Уклонение от налогов приводит к снижению поступления налогов, выступающих основным источником формирования доходной части бюджета, что в свою очередь негативно влияет на экономическое состояние государства в целом.

Применение Data Mining поможет систематизировать и установить субъекты, основания, предмет, форму и методы документальных проверок, проводимых налоговой полицией; уяснить сущность и специфику объектов, подлежащих исследованию в ходе проверок.

Для начала рассмотрим, организацию процесса Data Mining и разработку Data Mining-систем. Data Mining является непрерывным процессом со многими циклами и обратными связями и включает следующие фазы:

1. Осмысление бизнеса.
2. Осмысление данных.
3. Подготовка данных.
4. Моделирование.
5. Оценка результатов .
6. Внедрение.

К этому набору фаз иногда добавляют седьмой шаг - Контроль, он заканчивает круг.

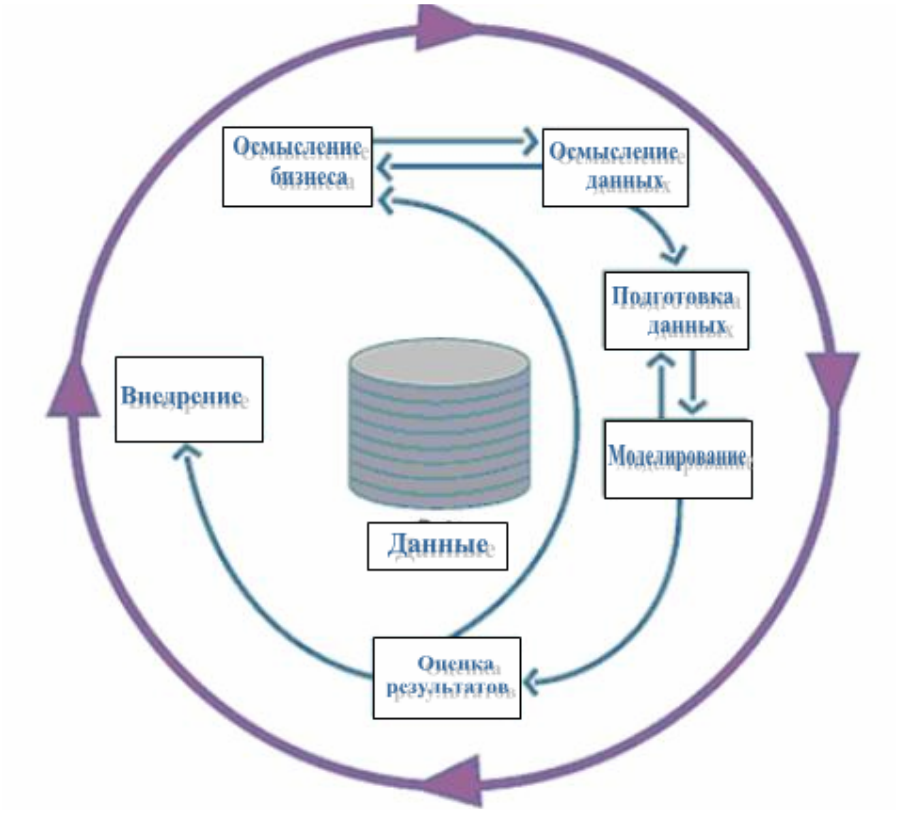

Рис. 4 Фазы процесса Data Mining

При помощи данной методологии Data Mining превращается в бизнес-процесс, в ходе которого технология Data Mining фокусируется на решении конкретных проблем. Эта методология разработана экспертами в индустрии Data Mining, представляет собой пошаговое руководство, где определены задачи и цели для каждого этапа процесса Data Mining.

Методология Data Mining описывается в терминах иерархического моделирования процесса, который состоит из набора задач, описанных четырьмя уровнями обобщения: фазы, общие задачи, специализированные задачи и запросы.

На верхнем уровне процесс Data Mining организовывается в определенное количество фаз, на втором уровне каждая фаза разделяется на несколько общих задач. Задачи второго уровня называются общими, потому что они являются обозначением (планированием) достаточно широких задач, которые охватывают все возможные Data Mining-ситуации. Третий уровень является уровнем специализации задачи, т.е. тем местом, где действия общих

задач переносятся на конкретные специфические ситуации. Четвертый уровень является отчетом по действиям, решениям и результатам фактического использования Data Mining.

CRISP-DM - это не единственный стандарт, описывающий методологию Data Mining. Помимо него, можно применять такие известные методологии, являющиеся мировыми стандартами, как Two Crows, SEMMA, а также методологии организации или свои собственные.

Исследование Data Mining основаны на понятии данных. В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки.

Иными словами, данные - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Форматы хранения данных. Одна из основных особенностей данных современного мира состоит в том, что их становится очень много. Возможны четыре аспекта работы с данными: определение данных, вычисление, манипулирование и обработка (сбор, передача и др.).

При манипулировании данными используется структура данных типа "файл". Файлы могут иметь различные форматы.

Большинство инструментов Data Mining позволяют импортировать данные из различных источников, а также экспортировать результирующие данные в различные форматы. Данные для экспериментов удобно хранить в каком-то одном формате.В некоторых инструментах Data Mining эти процедуры называются импорт/экспорт данных, другие позволяют напрямую открывать различные источники данных и сохранять результаты Data Mining в одном из предложенных форматов.

Наиболее распространенные форматы, согласно опросу "Форматы хранения данных", представлены на рисунке.

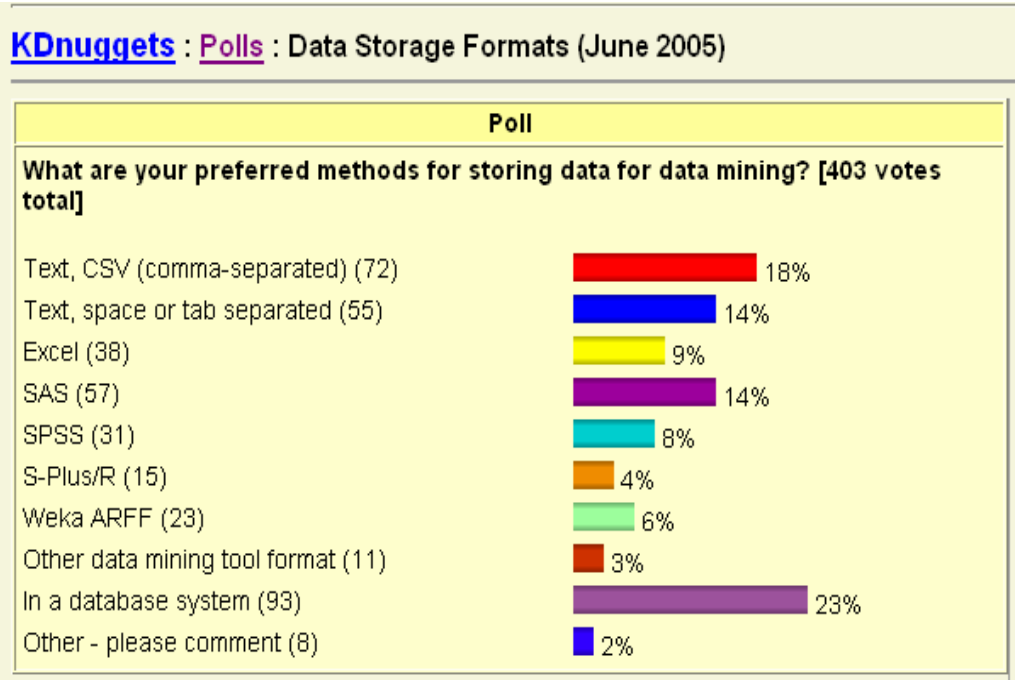

Рис. 5 Наиболее распространенные форматы хранения данных

Таким образом, наиболее распространенным форматом хранения данных для Data Mining выступают базы данных.

База данных (Database) - это особым образом организованные и хранимые в электронном виде данные.

Особым образом организованные означает, что данные организованы неким конкретным способом, способным облегчить их поиск и доступ к ним для одного или нескольких приложений. Также такая организация данных предусматривает наличие минимальной избыточности данных.

Базы данных являются одной из разновидностей информационных технологий, а также формой хранения данных.

Целью создания баз данных является построение такой системы данных, которая бы не зависела от программного обеспечения, применяемых технических средств и физического расположения данных в ЭВМ. Построение такой системы данных должно обеспечивать непротиворечивую и целостную информацию. При проектировании базы данных предполагается многоцелевое ее использование.

База данных в простейшем случае представляется в виде системы двумерных таблиц.Для управления базой данных необходима система управления базой данных (СУБД), которая представляет собой оболочку, с

помощью которой при организации структуры таблиц и заполнения их данными получается та или иная база данных.

Программные средства включают систему управления, обеспечивающую ввод-вывод, обработку и хранение информации, создание, модификацию и тестирование базы данных. Внутренними языками программирования СУБД являются языки четвертого поколения (C, C++, Pascal, Object Pascal,SQL). С помощью языков БД создаются приложения, базы данных и интерфейс пользователя, включающий экранные формы, меню, отчеты.

К базам данных, а также к СУБД предъявляются такие требования:

1. высокое быстродействие;
2. простота обновления данных;
3. независимость данных;
4. возможность многопользовательского использования *данных*;
5. безопасность данных;
6. стандартизация построения и эксплуатации БД (фактически СУБД);
7. адекватность отображения *данных* соответствующей предметной области;
8. дружелюбный интерфейс пользователя.

1.5 Методы и стадии Data Mining

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы; методы поиска ассоциативных правил, в том числе алгоритм Apriori; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов.

Большинство аналитических методов, используемые в технологии Data Mining - это известные математические алгоритмы и методы. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств. Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта.

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Понятие алгоритма появилось задолго до создания электронных вычислительных машин. Сейчас алгоритмы являются основой для решения многих прикладных и теоретических задач в различных сферах человеческой деятельности, в большинстве - это задачи, решение которых предусмотрено с использованием компьютера.

Алгоритм - точное предписание относительно последовательности действий, преобразующих исходные данные в искомый результат.

Data Mining может состоять из двух или трех стадий:

Стадия 1. Выявление закономерностей (свободный поиск).

Стадия 2. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование).

В дополнение к этим стадиям иногда вводят стадию валидации, следующую за стадией свободного поиска. Цель валидации - проверка достоверности найденных закономерностей. Однако, можно считать валидацию частью первой стадии, поскольку в реализации многих методов, в частности, нейронных сетей и деревьев решений, предусмотрено деление общего множества данных на обучающее и проверочное, и последнее позволяет проверять достоверность полученных результатов.

Стадия 3. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

Итак, процесс Data Mining может быть представлен рядом таких последовательных стадий.

СВОБОДНЫЙ ПОИСК ->

-> ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ->

-> АНАЛИЗ ИСКЛЮЧЕНИЙ

1. Свободный поиск

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей. Предварительные гипотезы относительно вида закономерностей здесь не определяются.

Закономерность - существенная и постоянно повторяющаяся взаимосвязь, определяющая этапы и формы процесса становления, развития различных явлений или процессов.

Система Data Mining на этой стадии определяет шаблоны, для получения которых в системах OLAP, например, аналитику налогового департамента необходимо обдумывать и создавать множество запросов.

Здесь же аналитик освобождается от такой работы - шаблоны ищет за него система. Особенно полезно применение данного подхода в сверхбольших базах данных, где уловить закономерность путем создания запросов достаточно

сложно, для этого требуется перепробовать множество разнообразных вариантов.

Свободный поиск представлен такими действиями:

- выявление закономерностей условной логики
- выявление закономерностей ассоциативной;
- выявление трендов и колебаний .

Допустим, имеется база с данными о декларации на прибыль. В случае самостоятельного задания запросов аналитик может получить приблизительно такие результаты: средний уровень прибыли по конкретному виду деятельности составляет равен 1200 условных единиц. В случае свободного поиска система сама ищет закономерности, необходимо лишь задать целевую переменную. В результате поиска закономерностей система сформирует набор логических правил "если ..., то ...".

2. Прогностическое моделирование (Predictive Modeling)

Вторая стадия Data Mining - прогностическое моделирование - использует результаты работы первой стадии. Здесь обнаруженные закономерности используются непосредственно для прогнозирования.

Прогностическое моделирование включает такие действия:

- предсказание неизвестных значений;
- прогнозирование развития процессов.

В процессе прогностического моделирования решаются задачи классификации и прогнозирования.

При решении задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью, к одному из известных, предопределенных классов на основании известных значений.

При решении задачи прогнозирования результаты первой стадии (определение тренда или колебаний) используются для предсказания неизвестных (пропущенных или же будущих) значений целевой переменной (переменных).

Продолжая рассмотренный пример первой стадии, можем сделать следующий вывод.

3. Анализ исключений

На третьей стадии Data Mining анализируются исключения или аномалии, выявленные в найденных закономерностях.

Действие, выполняемое на этой стадии, - выявление отклонений. Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии свободного поиска.

Классификация технологических методов Data Mining

Все методы Data Mining подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

1. Непосредственное использование данных, или сохранение данных.

В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. Выявление и использование формализованных закономерностей, или дистилляция шаблонов.

При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода Data Mining. Этот процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует. На стадиях прогностического моделирования *и* анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных. Напомним, что конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

Методы этой группы: логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях.

Методы Data Mining также можно классифицировать по задачам Data Mining. В соответствии с такой классификацией выделяем две группы. Первая из них - это подразделение методов Data Mining на решающие задачи

сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования.В соответствии со второй классификацией по задачам методы Data Mining могут быть направлены на получение описательных и прогнозирующих результатов. Описательные методы служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные методы кластерного анализа, в том числе: алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации и другие.

Прогнозирующие методы используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие методы: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов.

Другая классификация разделяет все многообразие методов Data Mining на две группы: статистические и кибернетические методы. Эта схема разделения основана на различных подходах к обучению математических моделей.Следует отметить, что существует два подхода отнесения статистических методов к Data Mining. Первый из них противопоставляет статистические методы и Data Mining, его сторонники считают классические статистические методы отдельным направлением анализа данных. Согласно второму подходу, статистические методы анализа являются частью математического инструментария Data Mining. Большинство авторитетных источников придерживается второго подхода.

В этой классификации различают две группы методов:

- статистические методы, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
- кибернетические методы, включающие множество разнородных математических подходов.

Недостаток такой классификации: и статистические, и кибернетические алгоритмы тем или иным образом опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к извлечению знаний из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах Data Mining.

Различные методы Data Mining характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств.

Среди основных свойств и характеристик методов Data Mining рассмотрим следующие: точность, масштабируемость, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

Масштабируемость - свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и пр., при добавлении к ней вычислительных ресурсов.

В таблице приведена сравнительная характеристика некоторых распространенных методов. Оценка каждой из характеристик проведена следующими категориями, в порядке возрастания: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/-высокая, высокая, очень высокая.

Как видно из рассмотренной таблицы, каждый из методов имеет свои сильные и слабые стороны. Но ни один метод, какой бы не была его оценка с точки зрения присущих ему характеристик, не может обеспечить решение всего спектра задач Data Mining.

Большинство инструментов Data Mining, предлагаемых сейчас на рынке программного обеспечения, реализуют сразу несколько методов, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, самоорганизующиеся карты Кохонена и визуализацию.

Сравнительная характеристика методов Data Mining								
Алгоритм	**Точность**	**Масштаби-руемость**	**Интерпрети-руемость**	**Пригод-ность к исполь-зованию**	**Трудо-емкость**	**Разносто-ронность**	**Быстрота**	**Популяр-ность, широта исполь-зования**
класси-ческие *методы* (линейная регрессия)	нейтраль-ная	высокая	высокая / нейтраль-ная	высокая	нейтраль-ная	нейтраль-ная	высокая	низкая
нейронные сети	высокая	низкая	низкая	низкая	нейтраль-ная	низкая	очень низкая	низкая
методы визуали-зации	высокая	очень низкая	высокая	высокая	очень высокая	низкая	чрезвы-чайно низкая	высокая / нейтраль-ная
деревья решений	низкая	высокая	высокая	высокая / нейтраль-ная	высокая	высокая	высокая / нейтраль-ная	высокая / нейтраль-ная
полино-миальные нейронные сети	высокая	нейтральная	низкая	высокая / нейтраль-ная	нейтраль-ная / низкая	нейтраль-ная	низкая / нейтраль-ная	нейтраль-ная
k-ближай-шего соседа	низкая	очень низкая	высокая / нейтраль-ная	нейтраль-ная	нейтраль-ная / низкая	низкая	высокая	низкая

Таблица 1. Сравнительная характеристика методов Data Mining

В универсальных прикладных статистических пакетах (например, SPSS, SAS, STATGRAPICS, Statistica, др.) реализуется широкий спектр разнообразнейших методов (как статистических, так и кибернетических). Следует учитывать, что для возможности их использования, а также для интерпретации результатов работы статистических методов (корреляционного, регрессионного, факторного, дисперсионного анализа и др.) требуются специальные знания в области статистики.

Универсальность того или иного инструмента часто накладывает определенные ограничения на его возможности. Преимуществом использования таких универсальных пакетов является возможность относительно легко сравнивать результаты построенных моделей, полученные различными методами. Такая возможность реализована, например, в пакете Statistica, где сравнение основано на так называемой "конкурентной оценке моделей". Эта оценка состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик для выбора наилучшей из них.

Большинство инструментов Data Mining, предлагаемых сейчас на рынке программного обеспечения, реализуют сразу несколько методов, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, самоорганизующиеся карты Кохонена и визуализацию.

В основу технологии Data Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих, скрытых от невооруженного глаза закономерностей решаются задачи Data Mining. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные задачи Data Mining.

Единого мнения относительно того, какие задачи следует относить к Data Mining, нет. Большинство авторитетных источников перечисляют следующие: классификация, кластеризация, прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов.

Классификация - наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Для решения задачи классификации могут использоваться методы: ближайшего соседа; k-ближайшего соседа; байесовские сети; индукция деревьев решений; нейронные сети .

Кластеризация - является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.

В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие **ассоциации** от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern).

Правило последовательности: после события X через определенное время произойдет событие Y.

Прогнозирование. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Определение отклонений или выбросов. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Оценивание .Задача оценивания сводится к предсказанию непрерывных значений признака.

Анализ связей - задача нахождения зависимостей в наборе данных.

Визуализация .В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

Пример методов визуализации - представление данных в 2-D и 3-D измерениях. Подведение итогов - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

Следует отметить, что уровни анализа (данные, информация, знания) практически соответствуют этапам эволюции анализа данных, которая происходила на протяжении последних лет.

Верхний - уровень приложений - является уровнем бизнеса (если мы имеем дело с задачей бизнеса), на нем менеджеры принимают решения.

Приведенные примеры приложений: перекрестные продажи, контроль качества, удерживание клиентов.

Средний - уровень действий - по своей сути является уровнем информации, именно на нем выполняются действия Data Mining; на рисунке приведены такие действия: прогностическое моделирование, анализ связей, сегментация данных и другие.

Нижний - уровень определения задачи Data Mining, которую необходимо решить применительно к данным, имеющимся в наличии; на рисунке приведены задачи предсказания числовых значений, классификация, кластеризация, ассоциация.

Рассмотрим таблицу, демонстрирующую связь этих понятий.

Уровни Data Mining				
уровень 3	приложения	выявление зон рисков налогоплательщиков	знания	Data Mining результат
уровень 2	действия	прогностическое моделирование	информация	метод анализа
уровень 1	задачи	классификация	данные	запросы

Таблица 2. Уровни Data Mining

Напомним, что для решения задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью, к одному из известных, предопределенных классов на основании известных значений.

Рассмотрим задачу удержания клиентов (определения надежности клиентов фирмы).

Первый уровень. Данные - база данных по налогоплательщикам. Есть данные о налогоплательщике (сфера деятельности, начисленные поступления, предъявленные декларации). Определенная часть налогоплательщиков, выполняют налоговые обязательства вовремя; другие налогоплательщики, вследствие небыполнения налоговых обязательств , приобретают задолженность перед государственным бюджетом. На этом уровне мы определяем тип задачи - это задача классификации.

На **втором уровне** определяем действие - прогностическое моделирование. С помощью прогностического моделирования мы с

определенной долей уверенности можем отнести объект, в данном случае, налогоплательщика, к одному из известных классов.

На **третьем уровне** мы можем воспользоваться приложением для принятия решения. В результате приобретения знаний, министрерство финансов может существенно снизить расходы, например, камеральные проверки, зная заранее, на какую организацию стоит обратить большее внимание.

Для получения ценных знаний необходимы качественные процедуры обработки. Процесс перехода от данных к знаниям занимает много времени и стоит дорого. Поэтому очевидно, что технология Data Mining с ее мощными и разнообразными алгоритмами является инструментом, при помощи которого, продвигаясь вверх по информационной пирамиде, мы можем получать действительно качественные и ценные знания**.**

Задачи Data Mining

Классификацией является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining.

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Различают:

- вспомогательную (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- естественную классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

Методы, применяемые для решения задач классификации

Для *классификации* используются различные методы. Основные из них:

- *классификация* с помощью деревьев решений;
- байесовская (наивная) *классификация*;
- *классификация* при помощи искусственных нейронных сетей;
- *классификация* методом опорных векторов;
- статистические методы, в частности, линейная регрессия;
- *классификация* при помощи метода ближайшего соседа;
- *классификация* CBR-методом;
- *классификация* при помощи генетических алгоритмов.

Схематическое решение задачи классификации некоторыми методами (при помощи линейной регрессии, деревьев решений и нейронных сетей) приведены рисунках

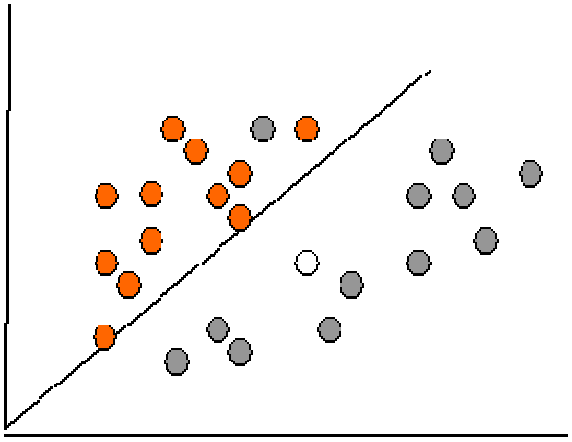

Рис. 6 Решение задачи классификации методом линейной регрессии

```
if X > 5 then grey
  else if Y > 3 then orange
    else if X > 2 then grey
      else orange
```

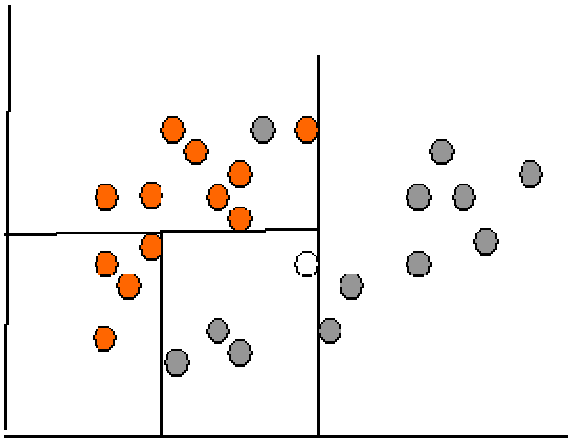

Рис.7 Решение задачи классификации методом девевьев решений

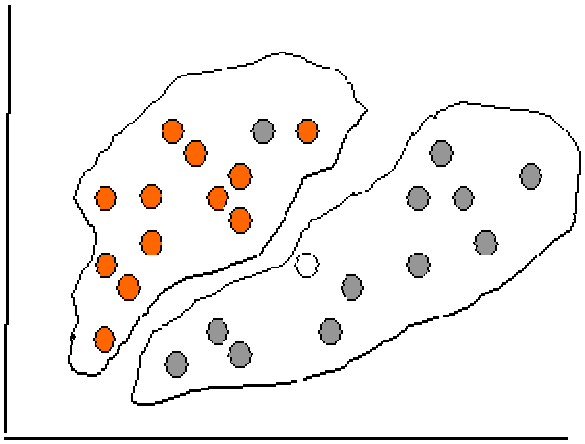

Рис. 8 Решение задачи классификации методом нейронных сетей

Оценивание классификационных методов

Оценивание методов следует проводить, исходя из следующих характеристик : скорость, робастность, интерпретируемость, надежность.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность, т.е. устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Свойства классификационных правил:

- размер дерева решений;

- компактность классификационных правил.

Надежность методов классификации предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

Прогнозирование и визуализация

Задача прогнозирования, пожалуй, может считаться одной из наиболее сложных задач Data Mining, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа.

Прогнозирование, в широком понимании этого слова, определяется как опережающее отражение будущего. Целью прогнозирования является предсказание будущих событий.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных.

Прогнозирование является распространенной и востребованной задачей во многих областях человеческой деятельности. В результате прогнозирования уменьшается риск принятия неверных, необоснованных или субъективных решений.

Примеры его задач: прогноз движения денежных средств, прогнозирование урожайности агрокультуры, прогнозирование финансовой устойчивости предприятия.

Визуализация - это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения. В результате использования визуализации создается графический образ данных. Применение визуализации помогает в процессе анализа данных увидеть аномалии, структуры. Визуализации данных может быть представлена в виде: графиков, схем, гистограмм, диаграмм и т.д.

Глава 2

Методология и реализация компьютерных моделей

2.1 Модели струкруры данных налогового департамента

В следствие каждодневных операций, производимых на личной карте каждого налогоплательщика, скапливается большой обьем информации, требующий систематизации. Эффективное хранение информации достигается наличием в составе информационно-аналитической системы целого ряда источников данных. Обработка и объединение информации достигается применением инструментов извлечения, преобразования и загрузки данных. Анализ данных осуществляется при помощи современных инструментов анализа данных.

Обычно в любой организации, в том числе и в министерсве финансов Грузии, функционирует несколько различных несвязанных или слабо связанных СОД, выгруженные из них данные, как правило, имеют различную структуру, формат, стандарты представления дат. Для обозначения одних и тех же объектов, используются различные кодировки. Как правило, в них, в явном виде отсутствуют реквизиты, идентифицирующие временной срез, которому они соответствуют и источники их получения.

В результате, огромные архивные массивы, накопленные за годы эксплуатации СОД и содержащие самую разнообразную жизненно важную для организации информацию, остаются невостребованными. Без предварительной доработки и согласования, архивные данные бесполезны и не могут быть непосредственно использованы в задачах анализа.

В основе концепции Хранилищ Данных лежат две основополагающие идеи:

- Интеграция ранее разъединенных детализированных данных:
    - исторические архивы,
    - данные из традиционных СОД,
    - данные из внешних источников в едином Хранилище Данных, их согласование и возможно агрегация.
- Разделение наборов данных используемых для операционной обработки и наборов данных используемых для решения задач анализа.

Поэтому для реализации хранилищ данных обычно используется несколько продуктов, одни их которых представляют собой собственно средства хранения данных, другие — средства их извлечения и просмотра, третьи — средства их пополнения и т.д.

Типичное хранилище данных, как правило, отличается от обычной реляционной базы данных. Во-первых, обычные базы данных предназначены для того, чтобы помочь пользователям выполнять повседневную работу, тогда как хранилища данных предназначены для принятия решений.[24]

Обычные базы данных подвержены постоянным изменениям в процессе работы пользователей, а хранилище данных относительно стабильно: данные в нем обычно обновляются согласно расписанию (например, еженедельно, ежедневно или ежечасно — в зависимости от потребностей). В идеале процесс пополнения представляет собой просто добавление новых данных за определенный период времени без изменения прежней информации, уже находящейся в хранилище. И в-третьих, обычные базы данных чаще всего являются источником данных, попадающих в хранилище.

Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных. Технология комплексного многомерного анализа данных получила название OLAP (On-Line Analytical Processing). OLAP — это ключевой компонент организации хранилищ данных. Концепция OLAP была описана в 1993 году Эдгаром Коддом. Коддом, был сформулирован так называемый тест FASMI, включающий следующие требования к приложениям для многомерного анализа:

- предоставление пользователю результатов анализа за приемлемое время (обычно не более 5 с), пусть даже ценой менее детального анализа;
- возможность осуществления любого логического и статистического анализа, характерного для данного приложения, и его сохранения в доступном для конечного пользователя виде;
- многопользовательский доступ к данным с поддержкой соответствующих механизмов блокировок и средств авторизованного доступа;
- многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий (это — ключевое требование OLAP);

возможность обращаться к любой нужной информации независимо от ее объема и места хранения.

В основе концепции хранилища данных лежат две основные идеи - интеграция разъединенных детализированных данных (детализированных в том смысле, что они описывают некоторые конкретные факты, свойства, события и т.д.) в едином хранилище и разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа. в период бурного развития регистрирующих информационных систем, возникло понимание ограниченности возможности их применения для целей анализа данных и построения на их основе систем поддержки и принятия решений. Регистрирующие системы создавались для автоматизации рутинных операций по ведению бизнеса – выписка счетов, оформление договоров, проверка состояния склада и т.д., и основными пользователями таких систем был линейный персонал. Основными требованиями к таким системам были обеспечение транзакционности вносимых изменений и максимизация скорости их выполнения. Именно эти требования определили выбор реляционных СУБД и модели представления данных «сущность-связь» в качестве основных используемых технических решений при построении регистрирующих систем.
Для менеджеров и аналитиков в свою очередь требовались системы, которые бы позволяли:

- Анализировать информацию во временном аспекте;
- Формировать произвольные запросы к системе;
- Обрабатывать большие объемы данных;
- Интегрировать данные из различных регистрирующих систем.

Очевидно, что регистрирующие системы не удовлетворяли ни одному из вышеуказанных требований. В регистрирующей системе информация актуальна только на момент обращения к базе данных, в следующий момент времени по тому же запросу можем получить совершенно другой результат. Возможность обработки больших массивов данных также мала из-за настройки СУБД на выполнение коротких транзакций и неизбежного замедления работы остальных пользователей. Ответом на возникшую потребность стало появление новой технологии организации баз данных – технологии хранилищ данных.

2.2 Повышение производительности в хранилищах данных и системах поддержки принятия решений с исползованием материальных представлений

2.2.1 Материализированные представления

Для повышения производительности в хранилищах данных и системах поддержки принятия решений используют материализованные представления, которые многократно ускоряют выполнение запросов, обращающихся к большому количеству (сотням тысяч или миллионам) записей. Говоря упрощенно, они позволяют за секунды (и даже доли секунд) выполнять запросы к терабайтам данных. Это достигается за счет прозрачного использования заранее вычисленных итоговых данных и результатов соединений таблиц.[23]

Предварительно вычисленные итоговые данные обычно имеют очень небольшой объем по сравнению с исходными данными.

Предположим, имеется база данных деклараций по налогу на прибыль, в которую загружены сведения о сотнях тысячах деклараций , и необходимо проанализировать начисления по регионам. Будут просмотрены все записи, данные агрегированы по регионам с выполнением необходимых вычислений. С помощью материализованного представления можно сохранить итоговые данные по регионам и обеспечить автоматическую поддержку этих данных системой. При наличии десяти регионов итоговые данные будут состоять из десяти записей, так что мы будем обращаться не к тысячам фактических записей, а только к десяти. Более того, при выполнении несколько измененного запроса, например по определенному региону, ответ на него тоже можно получить по этому материализованному представлению.

Ниже приведена статисика запроса до использования материализованного представления.

```
SET autotrace ON
SET timing ON
SELECT
        P.Tax_Year,SUBSTR(S.Partition_Id,1,1),
        COUNT(P.Doc_Mos_Nom) Tp_Count,
```

```
TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.L
abor,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F16,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.B
ad_Debt,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F17,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.I
nterest,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F18,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.R
esearch,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F19,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.D
epreciation,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F20,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.I
nsurance,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F21,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.R
epair,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F22,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.O
ther_Deduct,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F23,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND(SUM(NVL(P.P
revious_Loss,0))/SUM(NVL(Aggregate_Inc,0))*100,2 ))) F27,

TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0)),0,NULL,ROUND((SUM(NVL(P.
Aggregate_Inc,0))-(SUM(NVL(P.Beg_Stock, 0)) +
SUM(NVL(P.Raw_Materials, 0)) -SUM( NVL(P.End_Stock, 0))))
        /(SUM(NVL(P.Aggregate_Inc,0)))*100,2 )))
F9,TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0))+
SUM(NVL(P.Other_Inc,0)),0,NULL,
        ROUND((SUM(NVL(P.Aggregate_Inc,0))+ SUM(NVL(P.Other_Inc,0))-
(SUM(NVL(P.Beg_Stock, 0)) + SUM(NVL(P.Raw_Materials, 0)) -SUM(
NVL(P.End_Stock, 0))))
        /(SUM(NVL(P.Aggregate_Inc,0))+ SUM(NVL(P.Other_Inc,0)))*100,2
))) F10,
        TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0))+
SUM(NVL(P.Other_Inc,0)),0,NULL,ROUND((SUM(DECODE(
        SIGN(NVL(P.Aggregate_Inc, 0) + NVL(P.Other_Inc, 0) -
(NVL(P.Beg_Stock, 0) + NVL(P.Raw_Materials, 0) - NVL(P.End_Stock, 0))
         - ( NVL(P.Labor, 0) + NVL(P.Bad_Debt, 0) + NVL(P.Interest,
0) + NVL(P.Research, 0) +
      NVL(P.Depreciation, 0) + NVL(P.Insurance, 0) + NVL(P.Repair, 0)
+ NVL(P.Other_Deduct, 0))) ,1,
        NVL(P.Aggregate_Inc, 0) + NVL(P.Other_Inc, 0) -
(NVL(P.Beg_Stock, 0) + NVL(P.Raw_Materials, 0) - NVL(P.End_Stock, 0))
         - ( NVL(P.Labor, 0) + NVL(P.Bad_Debt, 0) + NVL(P.Interest,
0) + NVL(P.Research, 0) +
      NVL(P.Depreciation, 0) + NVL(P.Insurance, 0) + NVL(P.Repair, 0)
+ NVL(P.Other_Deduct, 0)),0,
        NVL(P.Aggregate_Inc, 0) + NVL(P.Other_Inc, 0) -
(NVL(P.Beg_Stock, 0) + NVL(P.Raw_Materials, 0) - NVL(P.End_Stock, 0))
         - ( NVL(P.Labor, 0) + NVL(P.Bad_Debt, 0) + NVL(P.Interest,
0) + NVL(P.Research, 0) +
      NVL(P.Depreciation, 0) + NVL(P.Insurance, 0) + NVL(P.Repair, 0)
+ NVL(P.Other_Deduct, 0)),0) - NVL(P.Previous_Loss,
0)))/(SUM(NVL(P.Aggregate_Inc,0))+ SUM(NVL(P.Other_Inc,0)))*100,2 )))
F11,
        TO_NUMBER(DECODE(SUM(NVL(P.Aggregate_Inc,0))+
SUM(NVL(P.Other_Inc,0)),0,NULL,
```

```
        ROUND(SUM(DECODE(SIGN(NVL(AGGREGATE_INC, 0) + NVL(OTHER_INC,
0) - (NVL(BEG_STOCK, 0) + NVL(RAW_MATERIALS, 0) - NVL(END_STOCK, 0))
         - ( NVL(LABOR, 0) + NVL(BAD_DEBT, 0) + NVL(INTEREST, 0) +
NVL(RESEARCH, 0) +
      NVL(DEPRECIATION, 0) + NVL(INSURANCE, 0) + NVL(REPAIR, 0) +
NVL(OTHER_DEDUCT, 0))) ,1,
        NVL(AGGREGATE_INC, 0) + NVL(OTHER_INC, 0) - (NVL(BEG_STOCK,
0) + NVL(RAW_MATERIALS, 0) - NVL(END_STOCK, 0))
         - ( NVL(LABOR, 0) + NVL(BAD_DEBT, 0) + NVL(INTEREST, 0) +
NVL(RESEARCH, 0) +
      NVL(DEPRECIATION, 0) + NVL(INSURANCE, 0) + NVL(REPAIR, 0) +
NVL(OTHER_DEDUCT, 0)),0,
        NVL(AGGREGATE_INC, 0) + NVL(OTHER_INC, 0) - (NVL(BEG_STOCK,
0) + NVL(RAW_MATERIALS, 0) - NVL(END_STOCK, 0))
         - ( NVL(LABOR, 0) + NVL(BAD_DEBT, 0) + NVL(INTEREST, 0) +
NVL(RESEARCH, 0) +
      NVL(DEPRECIATION, 0) + NVL(INSURANCE, 0) + NVL(REPAIR, 0) +
NVL(OTHER_DEDUCT, 0)),0
        ))-SUM(NVL(P.Loss,0))/(SUM(NVL(P.Aggregate_Inc,0))+
SUM(NVL(P.Other_Inc,0)))*100,2 ))) F12,

TO_NUMBER(DECODE(SUM(B.Assets),0,NULL,ROUND(((SUM(NVL(P.Aggregate_Inc
,0))+ SUM(NVL(P.Other_Inc,0)))/SUM(B.Assets))*100,2 ))) F13,
        TO_NUMBER(DECODE(((SUM(NVL(P.BEG_STOCK,0)) +
SUM(NVL(P.END_STOCK,0)))/2),0,NULL,
        ROUND(((SUM(NVL(P.BEG_STOCK,0)) + SUM(NVL(P.RAW_MATERIALS,0))
- SUM(NVL(P.END_STOCK,0)))
        /((SUM(NVL(P.BEG_STOCK,0)) +
SUM(NVL(P.END_STOCK,0)))/2))*100,2 ))) F14,
        SUM(NVL(P.Aggregate_Inc,0))+ SUM(NVL(P.Other_Inc,0)) F15
FROM Sts.PROFITS_DEC P ,
(SELECT Doc_Mos_Nom,(SUM(NVL(Prev_End_Bal,0)) + SUM(NVL(End_Bal,0)))
/2 Assets
            FROM Sts.PROFITS_DEC_B
            GROUP BY Doc_Mos_Nom) B,
            sts.SACMIANO S
       WHERE
                   P.Sacm_Sache = S.Sacm_Sache
               AND P.Doc_Mos_Nom = B.Doc_Mos_Nom(+)

GROUP BY SUBSTR(S.Partition_Id,1,1),p.Tax_Year
ORDER BY P.Tax_Year,SUBSTR(S.Partition_Id,1,1);


------------------- --------
               4371 001
               . . . .
               7949 095
106 rows selected
```
Elapsed: 00:00:03.35

```
Statistics
----------------------------------------------------------
        195 recursive calls
          0 db block gets
       6498 consistent gets
        118 physical reads
        156 redo size
       2272 bytes sent via SQL*Net to client
       1066 bytes received via SQL*Net from client
         12 SQL*Net roundtrips to/from client
          3 sorts (memory)
```

```
          0 sorts (disk)

1     SELECT STATEMENT Optimizer Mode=CHOOSE(Cost=593, Card=74,
Bytes=2 K)
2       SORT GROUP BY
3         HASH JOIN
4           INDEX FAST FULL SCAN OF 'STS.PK_PRD'
4           TABLE ACCESS FULL OF 'STS.ORG'
```

Для получения результатов агрегирования необходимо просмотреть более 80000 записей в более чем 6498 блоках. Создав материализованное представление данных,можно избежать многократного подсчета по исходной таблице.

```
GRANT QUERY REWRITE TO ss;
ALTER SESSION SET query_rewrite_enabled=TRUE;
ALTER SESSION SET query_rewrite_integrity=enforced;
CREATE MATERIALIZED VIEW m_prof
 BUILD IMMEDIATE
 REFRESH ON COMMIT
 ENABLE QUERY REWRITE
 AS
 SELECT COUNT(p.sa_id_no), p.ins
          FROM PROF p
GROUP BY p.ins
Materialized view created.
analyze table m_prof compute statistics;
Table analyzed.
```

Elapsed: 00:00:03.35

```
Statistics
----------------------------------------------------------
        195 recursive calls
          0 db block gets
         13 consistent gets
        118 physical reads
        156 redo size
       2272 bytes sent via SQL*Net to client
       1066 bytes received via SQL*Net from client
         12 SQL*Net roundtrips to/from client
          3 sorts (memory)
          0 sorts (disk)

1     SELECT STATEMENT Optimizer Mode=CHOOSE(Cost=593, Card=74,
Bytes=2 K)
2       SORT GROUP BY
3         HASH JOIN
4           INDEX FAST FULL SCAN OF 'STS.PK_PRD'
4           TABLE ACCESS FULL OF 'STS.ORG'
```

По сути, мы заранее вычислили количество объектов и задали итоговую информацию в виде материализованного представления. Мы потребовали немедленно построить и наполнить данными это представление.

Вместо более чем 6498 consistent gets (логических операций ввода-вывода) использовано всего 106. Физического ввода-вывода на этот раз вообще не было данные взяты из кэша. Теперь буферный кэш будет значительно эффективнее, так как кэшировать надо намного меньше данных. Раньше кэширование рабочего множества даже не начиналось, но теперь все рабочее множество помещается в кэше.

При получении запроса

```
SELECT P.Tax_Year,SUBSTR(S.Partition_Id,1,1),
COUNT(P.Doc_Mos_Nom) Tp_Count…
```

сервер автоматически направляет его к соответствующему материализованному представлению.

Предпосылки использования материализованных представлений можно сформулировать коротко:

повышение производительности. Получив однажды ответы на сложные вопросы, можно существенно снизить нагрузку на сервер.

При этом:

• **Уменьшается количество физических чтений.** Приходится просматривать меньше данных.

• **Уменьшается количество записей.** Не нужно так часто сортировать/агрегировать данные.

• **Уменьшается нагрузка на процессор.** Не придется постоянно вычислять агрегаты и функции от данных, поскольку это уже сделано.

• **Существенно сокращается время ответа.** При использовании итоговых данных запросы выполняются значительно быстрее по сравнению с запросами к исходным

данным. Все зависит от объема действий, которых можно избежать при использовании материализованного представления, но ускорение на несколько порядков вполне возможно.

При использовании материализованных представлений увеличивается потребность только в одном ресурсе — дисковом пространстве. Необходимо дополнительное место для хранения материализованных представлений, но за счет этого можно получить много преимуществ.

Материализованные представления больше подходят для сред, где данные только читаются. Они не предназначены для использования в среде интенсивной обработки транзакций. Они требуют дополнительных затрат

ресурсов при изменении базовых таблиц для учета этих изменений. Самое главное, что его использование полностью прозрачно для приложения и пользователя. Не нужно сообщать пользователям, какие итоговые таблицы поддерживаются, об этом информируется сервер Oracle с помощью требований целостности ссылок и измерений. Все остальное сервер сделает автоматически.

Таким образом материализированные представления помогают нам на этапе подготовки данных, сокращают время выполнения запроса, при создании разничных шаблонов, определающимися критериами риска, что в значительной мере влияет на эффективность работы программы моделей оценки рисков налогоплательщиков.[71]

2.2.2 Фрагментация таблиц в Базе Данных ORACLE

Фрагментация - одно из мгогих средств повышения производительности в базах данных, которое многократно упрощяет управление таблицами.

Возможность фрагментации, то есть разбиения таблицы или индекса на несколько меньших, проще управляемых частей, впервые появилась в сервере Oracle версии 8.0. Логически для обращающегося к базе данных приложения есть только одна таблица или индекс. Физически же эта таблица или индекс могут состоять из многих десятков фрагментов. Каждый фрагмент — самостоятельный объект, с которым можно работать отдельно или как с частью большего объекта. [11]

Фрагментация разрабатывалась для упрощения управления очень большими таблицами и индексами за счет применения подхода "разделяй и властвуй". Предположим, в базе данных имеется индекс размером 10 Гбайт. Если необходимо перестроить этот нефрагментированный индекс, придется перестраивать весь индекс в один прием. Хотя сервер способен перестраивать такие индексы динамически, объем ресурсов, необходимых для полного пересоздания всего индекса размером 10 Гбайт, огромен. Потребуется еще не менее 10 Гбайт свободного пространства для хранения обоих экземпляров индекса, необходима временная таблица журнала транзакций для записи изменений, сделанных в базовой таблице за время пересоздания индекса, и т.д. С другой стороны, если индекс разбит на десять фрагментов размером 1 Гбайт, можно пересоздавать каждый фрагмент индекса отдельно, по одному. При этом потребуется только 10 процентов свободного пространства, которое понадобилось бы при пересоздании нефрагментированного индекса. Пересоздание индекса пройдет намного быстрее (возможно, раз в десять),намного уменьшится объем выполненных транзакциями изменений, которые придется учесть в новом индексе, и т.д.

Фрагментация может сделать устрашающие по поглощению ресурсов, а иногда даже невозможные в большой базе данных операции настолько же простыми, как в маленькой.

Для использования фрагментации имеются три причины:

- повышение доступности данных;
- упрощение администрирования;
- повышение производительности запросов и операторов ЯМД.

Фрагменты повышают доступность также благодаря сокращению времени простоя. Например, если таблица размером 100 Гбайт разбита на 50 фрагментов размером 2 Гбайт, восстановление в случае ошибок выполняется в 50 раз быстрее. Если один из фрагментов размером 2 Гбайта поврежден, для его восстановления нужно намного меньше времени, чем для восстановления таблицы размером 100 Гбайт. Таким образом, доступность повышается по двум направлениям: многие пользователи могут вообще не заметить, что данные были недоступны, благодаря тому, что сбойный фрагмент пропускается, а время простоя при сбое сокращается вследствие существенно меньшего объема работы,
необходимой для восстановления.

2.2.3 Упрощение администрирования

Упрощение администрирования связано с тем, что операции с маленькими объектами выполнять гораздо проще, быстрее и при этом требуется меньше ресурсов, чем в случае больших объектов. Например, если оказалось, что 50 процентов строк в таблице фрагментированы, и необходимо это исправить, фрагментация таблицы очень пригодится. Чтобы исключить фрагментацию строк, как правило, пересоздается объект, в данном случае — таблица. Если таблица размером 100 Гбайт, придется выполнять эту операцию одним большим "куском", последовательно, с помощью оператора **ALTER TABLE MOVE.** Если же эта таблица разбита на 25 фрагментов размером 4 Гбайта, можно перестраивать фрагменты по одному. Более того, если это делается в период минимальной загруженности сервера, можно даже выполнять операторы **ALTER TABLE MOVE** параллельно в отдельных сеансах, что позволяет сократить время пересоздания. Практически все, что делается с нефрагментированным объектом, можно сделать и с частью фрагментированного объекта.[13]

Еще один фактор, который необходимо учитывать при оценке влияния фрагментации на администрирование, — использование смещающегося окна данных в хранилищах данных и при архивировании. Во многих случаях необходимо сохранять доступными последние (по времени создания) N групп данных. Например, необходимо предоставлять данные за последние 12 месяцев или пять лет. При отсутствии фрагментации это обычно связано с множественной вставкой новых данных, а затем множественным удалением устаревших. Для этого необходимо выполнить множество операторов ЯМД, сгенерировать множество данных повторного выполнения и отката. При использовании фрагментации можно:

- загрузить в отдельную таблицу данные за новый месяц (или год, или любой другой период);
- полностью проиндексировать таблицу (эти шаги можно сделать вообще в другом экземпляре, а результаты перенести в текущую базу данных);
- добавить ее в конец фрагментированной таблицы;

• удалить самый старый фрагмент с другого конца фрагментированной таблицы.

Повышение производительности операторов ЯМД и запросов

Еще одним преимуществом фрагментации является повышение производительности запросов и операторов ЯМД. Повышение производительности операторов ЯМД связано с потенциальной возможностью распараллеливания. При распараллеливании операторов ЯМД сервер Oracle использует несколько потоков или процессов для выполнения операторов **INSERT**,**UPDATE** или **DELETE**. На многопроцессорной машине с большой пропускной способностью ввода-вывода потенциальное ускорение для операторов ЯМД, выполняющих множественные изменения, может быть весьма большим. В отличие от параллельных запросов (обработки несколькими процессами/потоками оператора **SELECT**), для распараллеливания операторов ЯМД требуется фрагментация (есть специальный случай параллельной непосредственной вставки, задаваемой с помощью подсказки / *+ **APPEND** */>
когда фрагментация не требуется). Если таблицы не фрагментированы, распараллелить операторы ЯМД не удастся. Сервер Oracle присваивает каждому объекту максимальную степень распараллеливания в зависимости от количества составляющих его фрагментов.

Схемы фрагментации таблиц

В настоящее время сервер Oracle поддерживает три способа фрагментации таблиц.

• **Фрагментация по диапазону.** Можно указать диапазоны значений данных, строки для которых должны храниться вместе. Например, все данные за январь 2001 года будут храниться в фрагменте 1, все данные за февраль 2001 года — в фрагменте 2 и т.д. Это, вероятно, самый популярный способ фрагментации в Oracle .

• **Фрагментация по хеш-функции.** К значению одного или нескольких столбцов применяется хешфункция, определяющая фрагмент, в который помещается строка.

• **Составная фрагментация.** Это сочетание фрагментации по диапазону и по хеш-функции. Можно сначала применить разбиение по диапазону значений

данных, а затем выбрать в пределах диапазона фрагмент на основе значения хеш-функции.

Следующий код и схемы наглядно демонстрируют применение этих способов фрагментации. Кроме того, операторы **CREATE TABLE** структурированы так, чтобы можно было понять синтаксис создания фрагментированной таблицы.

Рассмотрим пример смешанной фрагментации, когда строки фрагментируются и по диапазону, и по хеш-функции. Здесь фрагментация по диапазону будет выполняться для одного набора столбцов, а фрагментация по хеш-функции — для другого. Вполне допустимо использовать одни и те же столбцы в обоих условиях фрагментации:

```
CREATE TABLE hash_example
2 (hash_key_column date,
3 data varchar2(20)
4 )
5 PARTITION BY HASH (hash_key_column)
6 (partition part_l tablespace p1,
7 partition part_2 tablespace p2
8 )
9 /
Table created.
CREATE TABLE composite_example
2 (range_key_column date,
3 hash_key_column int,
4 data varchar2(20)
5 )
6 PARTITION BY RANGE (range_key_column)
7 subpartition by hash(hash_key_column) subpartitions 2
8 (
9 PARTITION part_l
10 VALUES LESS THAN(to_date('01-jan-2007','dd-mon-yyyy'))
11 (subpartition part_l_sub_1,
12 subpartition part_l_sub_2
13 ),
14 PARTITION part_2
15 VALUES LESS THAN(to_date('01-jan-2008','dd-mon-yyyy'))
16 (subpartition part_2_sub_1,
17 subpartition part_2_sub_2
18 )
19 )
20 /
Table created.
```

При смешанной фрагментации сервер Oracle сначала применяет правила фрагментации по диапазону, чтобы понять, к какому диапазону относятся данные, а затем - хеш-функцию, которая и определяет, в какой физический фрагмент попадет строка. Фрагментация по диапазону используется, когда данные логически разделяются по значениям. Классический пример — данные, привязанные к периоду времени. Фрагментация по кварталам, по финансовым годам, по месяцам. Фрагментация по диапазону во многих случаях позволяет пропускать фрагменты, в том числе для условий строгого равенства и условий, задающих диапазоны: меньше, больше, в указанных пределах и т.д. Фрагментация по хеш-функции подходит для данных, в которых не удается выделить естественные диапазоны значений, подходящие для фрагментации.

Фрагментация особенно полезна как средство повышения масштабируемости при увеличении размеров больших объектов в базе данных. Повышение же масштабируемости положительно сказывается на производительности, доступности данных и упрощает администрирование. Все три последствия крайне важны для разных категорий пользователей.

Глава 3
Разработка моделей системы планирования выездных налоговых проверок

3 .1 Цели разработки моделей планирования проверок

Определено, что государственная налоговая политика должна формироваться исходя из необходимости стимулирования позитивных структурных изменений в экономике, последовательного снижения совокупной налоговой нагрузки, качественного улучшения налогового администрирования.

Проводимое государством все последние годы облегчение налогового бремени путем снижения налоговых ставок, отмены отдельных налогов и снятия неоправданных ограничений создает оптимальные условия для ведения бизнеса и исполнения налоговых обязательств.

Качественное налоговое администрирование является одним из условий эффективного функционирования налоговой системы и экономики государства. Позитивное развитие основных составляющих налоговой политики государства, которыми являются снижение совокупной налоговой нагрузки и улучшение налогового администрирования, неразрывно связано с налоговым контролем, целью которого является обеспечение своевременного и полного поступления налогов и других обязательных платежей в бюджет, в том числе за счет достижения высокого уровня налоговой дисциплины и грамотности налогоплательщиков.

Основной и наиболее эффективной формой налогового контроля являются выездные налоговые проверки. В результате проведения выездных налоговых проверок налоговыми органами должны одновременно решаться несколько задач, наиболее важные из которых:

- выявление и пресечение нарушений законодательства о налогах и сборах;
- предупреждение налоговых правонарушений.

При этом выездные налоговые проверки должны отвечать требованиям безусловного обеспечения законных интересов государства и прав налогоплательщиков, повышения их защищенности от неправомерных

требований налоговых органов и создания для налогоплательщика максимально комфортных условий для исчисления и уплаты налогов.

В целях эффективного решения всех этих задач подготовлена концепция системы планирования выездных налоговых проверок (далее - концепция), предусматривающая новый подход к построению системы отбора объектов для проведения выездных налоговых проверок.

Согласно концепции планирование выездных налоговых проверок - это открытый процесс, построенный на отборе налогоплательщиков для проведения выездных налоговых проверок по критериям риска совершения налогового правонарушения, в том числе общедоступным. Ранее планирование выездных налоговых проверок являлось сугубо внутренней конфиденциальной процедурой налоговых органов.

В целях обеспечения системного подхода к отбору объектов для проведения выездных налоговых проверок, Концепция определяет алгоритм такого отбора. Отбор основан на качественном и всестороннем анализе всей информации, которой располагают налоговые органы (в том числе из внешних источников) и определении на ее основе «зон риска» совершения налоговых правонарушений.

Таким образом, в настоящей концепции планирование выездных налоговых проверок взаимоувязано с формированием и развитием у налогоплательщиков правильного понимания законодательства о налогах и сборах, убеждения в недопустимости его нарушения и необходимости точного соблюдения законов.

3.2 Содержание и виды налогового риска

Формулируя понятие «налоговый риск», необходимо подразумевать его негативный характер. Причем негативный характер налогового риска имеет определенные формы проявления не только для налогоплательщиков, но и для всех субъектов налоговых правоотношений.

Необходимо разграничивать понятие «налогового риска» для налогоплательщиков, налоговых агентов и других субъектов налоговых правоотношений, представляющих интересы государства.

Причем для каждого из них он будет иметь различные формы проявления. С учетом вышесказанного содержание понятия «налогового риска» может быть сформулировано следующим образом. Под налоговым риском понимается пасность для субъекта налоговых правоотношений понести финансовые и иные потери, связанные с процессом налогообложения, вследствие негативных отклонений для данного субъекта от предполагаемых им, основанных на действующих нормах права, состояниях будущего, из расчета которых им принимаются решения в настоящем. Данное определение подразумевает существование налогового риска не только для налогоплательщиков, но и для других участников налоговых правоотношений. Например, для государства в лице государственных органов исполнительной власти налоговый риск состоит в снижении поступления налогов, выступающих основным источником формирования доходной части бюджета.

Основными характеристиками налогового риска являются:

- связан с неопределенностью экономической и правовой информации;
- является неотъемлемой составляющей финансового риска;
- распространяется на участников налоговых правоотношений: налогоплательщиков, налоговых агентов и других субъектов;
- имеет негативный характер для всех участников налоговых правоотношений;
- проявляется для каждого участника налоговых правоотношений по-разному.

3.3 Структура отбора налогоплательщиков для проведения выездных налоговых проверок

Обоснованный выбор объектов для проведения выездных налоговых проверок невозможен без всестороннего анализа всей информации, поступающей в налоговые органы из внутренних и внешних источников.

К информации из внутренних источников относится информация о налогоплательщиках, полученная налоговыми органами самостоятельно в процессе выполнения ими функций, возложенных на налоговую службу.

К информации из внешних источников относится информация о налогоплательщиках, полученная налоговыми органами в соответствии с действующим законодательством или на основании соглашений по обмену информацией с контролирующими и правоохранительными органами, органами государственной власти и местного самоуправления, а также иная информация, в том числе общедоступная.

Проводимый с целью отбора налогоплательщиков для проведения выездных налоговых проверок анализ финансово-экономических показателей их деятельности содержит несколько уровней, в том числе:

- анализ сумм исчисленных налоговых платежей и их динамики, который позволяет выявить налогоплательщиков, у которых уменьшаются суммы начислений налоговых платежей;
- анализ сумм уплаченных налоговых платежей и их динамики, проводимый по каждому виду налога (сбора) с целью контроля за полнотой и своевременностью перечисления налоговых платежей;
- анализ показателей налоговой и (или) бухгалтерской отчетности налогоплательщиков, позволяющий определить значительные отклонения показателей финансово-хозяйственной деятельности текущего периода от аналогичных показателей за предыдущие периоды или же отклонения от среднестатистических показателей отчетности аналогичных хозяйствующих субъектов за определенный промежуток времени, а также выявить противоречия между сведениями, содержащимися в представленных документах, и (или) несоответствие информации, которой располагает налоговый орган;

- анализ факторов и причин, влияющих на формирование налоговой базы.

В случае выбора объекта для проведения выездной налоговой проверки налоговый орган определяет целесообразность проведения выездных налоговых проверок контрагентов и (или) аффилированных лиц проверяемого налогоплательщика.

В соответствии с основными целями и принципами настоящей Концепции выбор объектов для проведения выездных налоговых проверок построен на целенаправленном отборе, тщательном и постоянно проводимом, всестороннем анализе всей имеющейся у налоговых органов информации о каждом объекте независимо от его формы собственности и сумм налоговых обязательств. При осуществлении планирования подлежат анализу все существенные аспекты, как отдельной сделки, так и деятельности налогоплательщика в целом.

Приоритетными для включения в план выездных налоговых проверок являются те налогоплательщики, в отношении которых у налогового органа имеются сведения об их участии в схемах ухода от налогообложения или схемах минимизации налоговых обязательств и (или) результаты проведенного анализа финансово-хозяйственной деятельности налогоплательщика свидетельствуют о предполагаемых налоговых правонарушениях.

3.4 Критерии оценки рисков для налогоплательщиков

1. Налоговая нагрузка у данного налогоплательщика ниже ее среднего уровня по хозяйствующим субъектам в конкретной отрасли (виду экономической деятельности).

2. Отражение в бухгалтерской или налоговой отчетности убытков на протяжении нескольких налоговых периодов.

3. Отражение в налоговой отчетности значительных сумм налоговых вычетов за определенный период.

4. Опережающий темп роста расходов над темпом роста доходов от реализации товаров (работ, услуг).

5. Выплата среднемесячной заработной платы на одного работника ниже среднего уровня по виду экономической деятельности.

6. Неоднократное приближение к предельному значению установленных Налоговым кодексом величин показателей, предоставляющих право применять налогоплательщикам специальные налоговые режимы.

7. Отражение индивидуальным предпринимателем суммы расхода, максимально приближенной к сумме его дохода, полученного за календарный год.

8. Построение финансово-хозяйственной деятельности на основе заключения договоров с контрагентами-перекупщиками или посредниками («цепочки контрагентов») без наличия разумных экономических или иных причин (деловой цели).

9. Непредставление налогоплательщиком пояснений на уведомление налогового органа о выявлении несоответствия показателей деятельности.

10. Неоднократное снятие с учета и постановка на учет в налоговых органах налогоплательщика в связи с изменением места нахождения («миграция» между налоговыми органами).

11. численные соотношения начисленной и выданной зарплаты

12. Значительное отклонение уровня рентабельности по данным бухгалтерского учета от уровня рентабельности для данной сферы деятельности по данным статистики.

3.5 Методические основы анализа налоговых рисков

Каким же образом оценить величину возможных потерь от налогового риска? Для ответа на этот вопрос необходимо рассмотреть существующие способы анализа рисков и выбрать те из них, которые могут быть использованы для оценки налоговых рисков.

При оценке риска анализируют две его составляющие: вероятность наступления и характер ущерба (рис. 2). Вероятность наступления риска может быть определена объективным или субъективным методом. Объективный метод определения вероятности основан на вычислении частоты, с которой происходит рисковое событие. Субъективный метод определения вероятности основан на использовании различных предположений: суждений оценивающего, его личного опыта, оценки эксперта и т. п. Когда вероятность определяется субъективно, то различными субъектами анализа может устанавливаться разное ее значение для одного и того же события. Определение характера ущерба даже в случае субъективной оценки носит основанное на предположениях стоимостное выражение.

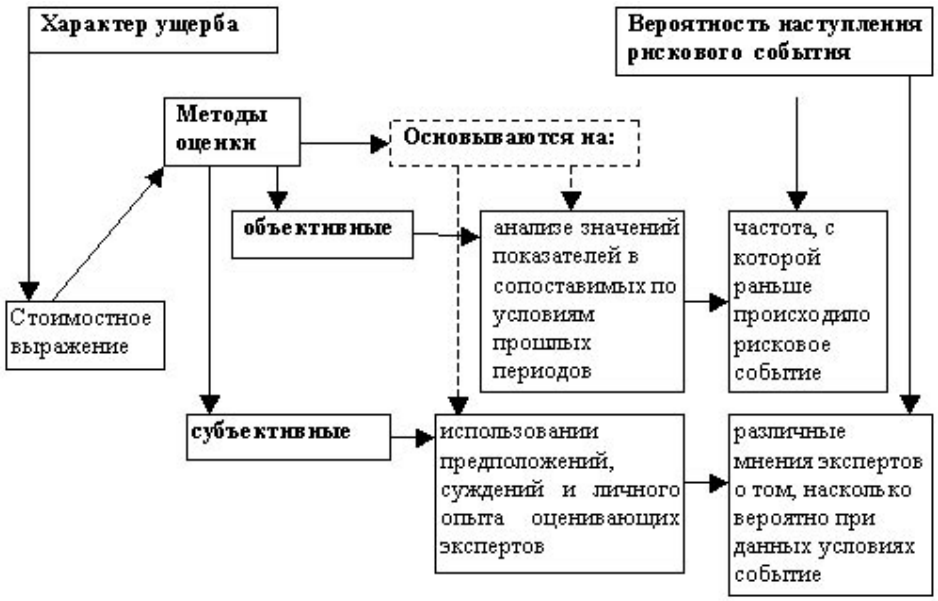

Рис.9 Анализируемые составляющие налоговых рисков

Для оценки вариантов налоговых решений по их рисковой составляющей аналитическую деятельность методически правильно организовать в виде следующих этапов:

1 этап - определение условий сравнения: целей налогового решения и задач, стоящих перед его оценкой; временного интервала (момента) оценки; требований к используемой налоговой и неналоговой информации и возможности их выполнения;

2 этап - формирование показателя - критерия сравнения необходимо осуществлять, руководствуясь определенными принципами, для реализации каждого из которых необходимо четко ответить на вопросы, приведенные в таблице;

Принципы формирования критерия оценки вариантов налоговых решений с учетом риска

Принципы формирования критерия	**Вопросы, на которые необходимо дать четкий ответ, формируя критерий оценки**
соответствие экономического содержания показателя целям оценки	содержание и место критерия в системе показателей финансового анализа
соответствие исходной информации требованиям	определяющие факторы и аналитическое выражение показателя
корректность построения показателя и рекомендуемых значений	возможность расчета критерия по сравниваемым вариантам

Таблица 3. Принципы формирования критерия

3 этап - расчет значений критерия по всем сравниваемым вариантам, их оценка и аналитическая интерпретация для принятия наиболее обоснованного решения.

3.6 Результаты внедрения системы планирования проверок

Настоящая система определяет основные приоритеты, принципы и направления реализации единого подхода к планированию выездных налоговых проверок.

Предлагаемая система планирования позволит:

1) для налогоплательщиков максимально уменьшить вероятность того, что выездная налоговая проверка в текущем году затронет законопослушного налогоплательщика, полностью исполнившего свои обязательства перед бюджетом;

2) для налоговых органов выявить наиболее вероятные «зоны риска» (нарушения законодательства о налогах и сборах), своевременно отреагировать на возможное совершение налоговых правонарушений и определить необходимые мероприятия налогового контроля.

Принципы, заложенные в настоящей системы, позволят реализовать:

1. Формирование единого подхода к планированию выездных налоговых проверок.
2. Стимулирование налогоплательщиков в части соблюдения законодательства о налогах и сборах.
3. Повышение налоговой грамотности и дисциплины налогоплательщиков.

Глава 4

Систематизация и анализ данных

4.1 Приемы систематизации данных и исследование моделей

На третьем этапе процесса Data Mining , т.е. подготовке данных, мы использовали такие инструменты оптимизации представления информациии, как хранилища данных, материализированные представления, фрагментация таблиц. Проведены процедуры анализа и модернизации с целью выявления ошибок, несоответствий и определения необходимых очищающих преобразований.

Применяя методы преобразования и очистки данных, позволяющие изменять вид представления, проводится нормализация значений, выявляются неопределенные или отсутствующие значения. На основе подготовленных данных специальные процедуры автоматически строят модели для дальнейшего прогнозирования, классификации новых ситуаций, выявления аналогий. Так же вычислены значения так называемого "среднего" .

Главная цель среднего - представление набора данных для последующего анализа, сопоставления и сравнения. Значение среднего легко вычисляется и может быть использовано для последующего анализа. Оно может быть вычислено для данных, измеряемых по интервальной шкале. Среднее значение рассчитывается как среднее арифметическое набора данных: сумма всех значений выборки, деленная на объем выборки. "Сжимая" данные таким образом, мы получаем информативный набор данных по соответствующему виду деятельноцти налогоплательщика.

Среднее значение очень информативно и позволяет делать вывод относительно всего исследуемого набора данных. При помощи среднего мы получаем возможность сравнивать несколько наборов данных или их частей.

Аналогично представлена информация о каждом налогоплательщике , предъявившем данную декларацию в материализированном представлении V_IND_TP. Следовательно, оперируя данными этих таблиц получим отклонение показателя налоговой нагрузки каждого налогоплательщика от среднего уровня по хозяйствующим субъектам в конкретной отрасли, что может

послужить сигналом о недостоверности представленной в декларации информации.

Во избежание потери временных и программых ресурсов, эти показатели записываем в базу данных в виде материализированного представления следующим образом, за счет чего многократно упрощяется написание программого кода, соответсевенно его читабельность, и значительно уменшается время считывания данных из базы:

```
CREATE OR REPLACE FORCE VIEW AUD.PROF_TP_VARIANCE
(TAX_YEAR, PARTITION_ID, SA_IDENT_NO, TP_COUNT, V_1,
 V_2, V_3, V_4, V_5, V_6,
 V_7, V_8, V_9, V_10, V_11,
 V_12, V_13, V_14, V_15, V_16)
AS
SELECT
         T.Tax_Year, T.Partition_Id, T.Sa_Ident_No, T.Tp_Count,

TO_NUMBER(DECODE(T.Col_1,NULL,NULL,DECODE(a.Col_1,0,NULL,ROUND
((T.Col_1 - A.Col_1)/a.Col_1,2)))) v_1,

TO_NUMBER(DECODE(T.Col_5,NULL,NULL,DECODE(a.Col_2,0,NULL,ROUND((T.Col
_2 - A.Col_2)/a.Col_2,2)))) v_2,

TO_NUMBER(DECODE(T.Col_3,NULL,NULL,DECODE(a.Col_3,0,NULL,ROUND((T.Col
_3 - A.Col_3)/a.Col_3,2)))) v_3,

TO_NUMBER(DECODE(T.Col_4,NULL,NULL,DECODE(a.Col_4,0,NULL,ROUND((T.Col
_4 - A.Col_4)/a.Col_4,2)))) v_4,

TO_NUMBER(DECODE(T.Col_5,NULL,NULL,DECODE(a.Col_5,0,NULL,ROUND((T.Col
_5 - A.Col_5)/a.Col_5,2)))) v_5,

TO_NUMBER(DECODE(T.Col_6,NULL,NULL,DECODE(a.Col_6,0,NULL,ROUND((T.Col
_6 - A.Col_6)/a.Col_6,2)))) v_6,

TO_NUMBER(DECODE(T.Col_7,NULL,NULL,DECODE(a.Col_7,0,NULL,ROUND((T.Col
_7 - A.Col_7)/a.Col_7,2)))) v_7,

TO_NUMBER(DECODE(T.Col_8,NULL,NULL,DECODE(a.Col_8,0,NULL,ROUND((T.Col
_8 - A.Col_8)/a.Col_8,2)))) v_8,

TO_NUMBER(DECODE(T.Col_9,NULL,NULL,DECODE(a.Col_9,0,NULL,ROUND((T.Col
_9 - A.Col_9)/a.Col_9,2)))) v_9,

TO_NUMBER(DECODE(T.Col_10,NULL,NULL,DECODE(a.Col_10,0,NULL,ROUND((T.C
ol_10 - A.Col_10)/a.Col_10,2)))) v_10,

TO_NUMBER(DECODE(T.Col_11,NULL,NULL,DECODE(a.Col_11,0,NULL,ROUND((T.C
ol_11 - A.Col_11)/a.Col_11,2)))) v_11,

TO_NUMBER(DECODE(T.Col_12,NULL,NULL,DECODE(a.Col_12,0,NULL,ROUND((T.C
ol_12 - A.Col_12)/a.Col_12,2)))) v_12,

TO_NUMBER(DECODE(T.Col_13,NULL,NULL,DECODE(a.Col_13,0,NULL,ROUND((T.C
ol_13 - A.Col_13)/a.Col_13,2)))) v_13,
```

```
TO_NUMBER(DECODE(T.Col_14,NULL,NULL,DECODE(a.Col_14,0,NULL,ROUND((T.C
ol_14 - A.Col_14)/a.Col_14,2)))) v_14,

TO_NUMBER(DECODE(T.Col_15,NULL,NULL,DECODE(a.Col_15,0,NULL,ROUND((T.C
ol_15 - A.Col_15)/a.Col_15,2)))) v_15,

TO_NUMBER(DECODE(T.Col_16,NULL,NULL,DECODE(a.Col_16,0,NULL,ROUND((T.C
ol_16 - A.Col_16)/a.Col_16,2)))) v_16

FROM Prof_Ind_Tp T, Prof_Ind_Avg A

WHERE
        T.Tax_Year = A.Tax_Year AND

        T.Partition_Id = A.Partition_Id
```

Теперь для получения данных из этой таблицы нам достаточно написать следующий код:

```
SELECT

        TAX_YEAR,
         PARTITION_ID,
        SA_IDENT_NO,
        TP_COUNT,
        V_16, V_17,
        V_18, V_19,
         V_20, V_21,
        V_22, V_23,
        V_9, V_10,
        V_11, V_12,
        V_13

FROM STS.PROF_TP_VARIANCE Vw

WHERE
        TAX_YEAR = :P_YEAR
  AND   PATRITION_ID := P_PART
```

Как видно , написание этого кода гораздо более просто, наглядно и удобочитаемо, а главное виполнение данного запроса не пребует больших временных затрат. Следовательно , использование материальных представлений способствует отимизации выполнения запросов , а так же их визуализации.

Итак, путем выборки данных из ранее созданных материальних представлений, получаем отклонение от среднего значения каждого пункта декларации на прибыль, в разрезе каждого вида экономической деятельности.

TAX_YEAR	PARTITION_ID	SA_IDENT_NO	TP_COUNT	V_1	V_2	V_3	V_4
2007	K	202171978	1.00	-0.64	-1.00	-1.00	-1.00
2007	A	230090482	1.00	3.66	-1.00	-1.00	-1.00
2007	B	204386952	1.00	-0.26	-1.00	-1.00	-1.00
2007	K	207066049	1.00	4.25	-1.00	-1.00	-1.00
2007	C	252642740	1.00	2.03	-1.20	1.80	-1.00

Таблица 4. Отклонение от среднего значения

Далее эти данные будут использоваться в процессе создания шаблонов при выявлении зон рисков налогоплательщиков.

4.2 Моделирование и анализ данных

Уделим внимание оставшимся этапам процесса Data Mining, а именно:

- построению модели;
- проверке и оценке моделей;
- выбору модели;
- применению модели;
- коррекции и обновлении модели.

Ключевым словом в названии всех этих этапов является понятие "модель".

Моделирование - единственный к настоящему времени систематизированный способ увидеть варианты будущего и определить потенциальные последствия альтернативных решений, что позволяет их объективно сравнивать. Моделирование широко применяется при использовании методов Data Mining. Путем использования моделей Data Mining осуществляется анализ данных. С помощи моделей Data Mining обнаруживается полезная, ранее неизвестная, доступная интерпретации информация, используемая для принятия решений. Создание и использование Data Mining модели является ключевым моментом для начала понимания, осмысления и прогнозирования тенденций анализируемого объекта.

Построение моделей Data Mining осуществляется с целью исследования или изучения моделируемого объекта, процесса, явления и получения новых знаний, необходимых для принятия решений. Использование моделей Data Mining позволяет определить наилучшее решение в конкретной ситуации.

Аналитик создает модель как подобие изучаемого объекта. Модели могут быть записаны в виде различных изображений, схем, математических формул и т.д. Преимуществом использования моделей при исследованиях является простота модели в сравнении с исследуемым объектом. При этом модели позволяют выделить в объекте наиболее существенные факторы с точки зрения цели исследования, и не отвлекаться на маловажные детали.Из последнего замечания следует, что модель обладает свойством неполноты, поскольку является по своему определению абстрактной.

Среди большого разнообразия методов Data Mining должен быть выбран метод или же комбинация методов, при использовании которых построенная модель будет наилучшим образом описывать исследуемый объект.

Иногда для выявления искомых закономерностей требуется использование нескольких методов и алгоритмов. В таком случае одни методы используются в начале моделирования, другие - на дальнейших этапах. Пример: для определения однотипных групп клиентов применялся один из методов кластеризации, в результате клиенты были разбиты на группы, каждой из которых, присвоен код; далее мы пользовались методом деревьев решений. Код группы (результат работы предыдущего метода) использовался для интерпретации полученных закономерностей.

Выбор метода, на основе которого будет построена модель, должен осуществляться с учетом постановки задачи, особенностей набора исходных данных, специфики решаемой задачи, результатов, которые должны быть получены на выходе.

Постановка задачи формализует суть задачи, так, наличие входных и выходных переменных при решении задачи классификации определяет выбор одного из методов "обучение с учителем". Наличие лишь входных переменных определяет выбор другого - метода "обучение без учителя".

Среди особенностей исходного набора данных, например, могут быть следующие его характеристики:

- количество записей в наборе;
- соотношение количества записей в наборе данных и количества входных переменных;
- наличие выбросов, ибо некоторые методы особенно чувствительны к наличию выбросов в данных. Этот факт следует учитывать при построении *модели* на подобных данных.

Как уже упоминалось выше, Data Mining является итеративным процессом.

Итерация - это циклическая управляющая структура, она содержит выбор между альтернативами и следование избранной.

Выбор между альтернативами в нашем случае - это этап оценки модели.

Если модель приемлема, возможно ее использование.

Этапы подготовки данных, построения модели, оценки модели и выбора лучшей представляют собой цикл.Если по каким-либо причинам построенная модель оказалось неприемлемой, цикл повторяется и следует один из следующих этапов:

- подготовка данных (если причина некорректности *модели* - в данных);

- построение модели (если причина некорректности - во внутренних параметрах самой модели).

Для определения специфических свойств исследуемых данных иногда требуется несколько итераций.

Цикл № t-1.

Подготовка данных -> построение модели № t-1-> оценка и выбор модели.

Цикл № t.

Подготовка данных -> построение модели № t -> оценка и выбор модели.

Цикл № t+1.

Подготовка данных -> построение модели № t+1 -> оценка и выбор модели.

Иногда имеет смысл использовать несколько методов параллельно для возможности сравнения и анализа данных с различных точек зрения.

Основные характеристики модели, которые определяют ее выбор, - это точность модели и эффективность работы алгоритма.

В некоторых программных продуктах реализован ряд методов, разработанных для выбора модели. Многие из них основаны на так называемой "конкурентной оценке моделей", которая состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик.

Например, в пакете Statistica (Statsoft) эти методы рассматриваются как ядро "предсказывающей добычи данных", они включают: накопление (голосование, усреднение); бустинг; мета-обучение.

После тестирования, оценки и выбора модели следует этап применения модели. На этом этапе выбранная модель используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining. Для классификационных и прогнозирующих моделей на этом этапе прогнозируется целевой (выходной) атрибут (target attribute).

По прошествии определенного установленного промежутка времени с момента начала использования модели Data Mining следует проанализировать полученные результаты, определить, действительно ли она "успешна" или же возникли проблемы и сложности в ее использовании.

Однако даже если модель с успехом используется, ее не следует считать абсолютно верной на все времена. Дело в том, что необходимо периодически оценивать адекватность модели набору данных, а также текущей ситуации

(следует учитывать возможность изменения внешних факторов). Даже самая точная *модель* со временем перестает быть таковой. Для того чтобы построенная модель выполняла свою функцию, следует работать над ее коррекцией (улучшением). При появлении новых данных требуется повторное обучение модели. Этот процесс называют обновлением модели. Работы, проводимые с моделью на этом этапе, также называют контролем и сопровождением модели.

Существует много причин, требующих обучить модель заново, т.е. обновить ее, чтобы отразить определенные изменения.

Основными причинами являются следующие:

- изменились входящие данные или их поведение;
- появились дополнительные данные для обучения;
- изменились требования к форме и количеству выходных данных;
- изменились цели бизнеса, которые повлияли на критерии принятия решений;
- изменилось внешнее окружение или среда (макроэкономика, политическая ситуация, научно-технический прогресс, появление новых конкурентов и товаров и т.д.).

Причины, перечисленные выше, могут обесценить допущения и исходную информацию, на которых основывалась *модель* при построении.

Таким образом, с помощью моделей решают задачи классификации и прогнозирования. Такое решение подразумевает двухэтапный процесс: создание модели и ее использование. При помощи классификационной модели решается следующая задача: выявление групп налогоплательщиков, попадающих в зону риска.

Программа выявления рисков позволяет обнаружить таких налаогоплательщиков, которые в силу своей финансовой деятельности требуют большего внимания со стороны налогового аудита. Для чего в процессе созданния модели учавствуют 12 идентификаторов, описанных выше.

Идентификаторы делятся на 2 группы.Впервой группе обьединены первые 2 идентификатора поскольку , в данном случае, необходимо сравнение конкретных показателей в зависимости от сферы экономическои деятельности со средним показателем. Во вторую группу входят все остальные идентификаторы, где прописываются конкретные значения.

На этапе построения модели при помощи классификационного метода и алгоритма была создана модель (классификатор налогоплательщиков).

Функции классификации предназначены для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект. Имеется столько же функций классификации, сколько групп. Каждая функция позволяет вам для каждого образца и для каждой совокупности вычислить веса классификации по формуле:

$$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + ... + w_{im}*x_m$$

В этой формуле индекс *i* обозначает соответствующую совокупность, а индексы *1, 2, ..., m* обозначают *m* переменных; c_i являются константами для *i*-ой совокупности, w_{ij} - веса для *j*-ой переменной при вычислении показателя классификации для *i*-ой совокупности; x_j - наблюдаемое значение для соответствующего образца *j*-ой переменной. Величина S_i является результатом показателя классификации.

Поэтому можно использовать функции классификации для прямого вычисления показателя классификации для некоторых новых значений.

Как только были бычислены показатели классификации для наблюдений, легко решить, как производить классификацию наблюдений. В общем случае наблюдение считается принадлежащим той совокупности, для которой получен наивысший показатель классификации. Поэтому, при изучении экономической активности налогоплательщика на основе нескольких переменных, наиболее вероятнее изпользовать функции классификации, чтобы предсказать, который из них вероятнее всего попадает в зону риска.

С одной стороны, можно говорить, что построенная модель выделила наиболее существенные (или значимые) факторы с точки зрения решаемой задачи. Для решения задачи классификации наиболее значимыми оказались переменные (идентификаторы) 5,7,8,9,11. Модели строятся автоматически на основе анализа имеющихся данных об объектах, наблюдениях и ситуациях с помощью специальных алгоритмов. Основу опции Data Mining составляют процедуры, реализующие алгоритмы построения моделей классификации. В нашем случае это хранимые процедуры заполняющие таблицы данными о налогоплательщиках , имеющих налоговую активность за время расчетного

периода, а так же рассчитиваюшие баллы , в зависимости от идентификатора и присвоенного коеффициента.

```
CREATE OR REPLACE PROCEDURE K_SCR_ORG_E IS
BEGIN

update AUD.K_ORG_AUD_SCOR_E set SCR= null;
update AUD.K_SUMS_CR set CR_SUM =null;

/

update AUD.K_SUMS_CR set CR_SUM =
(select CR1*k1+CR2*k2from (select
sum(decode(CRIT_ID,1,KOEF*ACTIVE,0)) k1,
            sum(decode(CRIT_ID,2,KOEF*ACTIVE,0)) k2      from
AUD.K_CRITERIAS_E));
COMMIT;
update AUD.K_ORG_AUD_SCOR_E set SCR =
(select aud.k_jg2_e(ID_CRIT, nvl(PRM_SXV,0))*nvl(KOEF,0)*ACTIVE
from AUD.K_CRITERIAS_E
where CRIT_ID=ID_CRIT)
where ID_CRIT in (1,2);
COMMIT;

/

update AUD.K_ORG_AUD_SCOR_E set SCR =
(select aud.k_jg2_e(ID_CRIT, nvl(PRM,0))*nvl(KOEF,0)*ACTIVE
from AUD.K_CRITERIAS_E
where CRIT_ID=ID_CRIT)
where ID_CRIT not in (1,2);
COMMIT;

END;
/


---

CREATE OR REPLACE PROCEDURE K_Cr_Scores IS
         ttt VARCHAR2(32000) DEFAULT '(SA_IDENT_NO,INS_KOD, RAI_KOD,
SACM_SACHE';
         ttt1 VARCHAR2(32000);
CURSOR x IS SELECT crit_ID, ID_TYPE_GR, CNT_GR, MIN_SCORE, MAX_SCORE,
KOEF
                  FROM AUD.K_CRITERIAS
                  WHERE ACTIVE=1
                  order by 1;
BEGIN
FOR cur IN x LOOP
      ttt := ttt || ', SC' || trim(TO_CHAR(cur.crit_id));

      ttt1 := ttt1 || ', ' || cur.koef*cur.max_score || '/' ||
(cur.cnt_gr - cur.MIN_SCORE);
IF cur.ID_TYPE_GR=1 THEN
      ttt1 := ttt1 || '*(CEIL(dense_rank() over (ORDER BY CR' ||
                  trim(TO_CHAR(cur.crit_id)) || ',SA_IDENT_NO)*' ||
                  TO_CHAR(cur.CNT_GR) || '/cnt) - ' || cur.MIN_SCORE
||
                  ') SC' || trim(TO_CHAR(cur.crit_id));
```

```
END IF;
IF cur.ID_TYPE_GR=2 THEN
      ttt1 := ttt1 || '*nvl((aud.k_jg2(' || cur.crit_id || ',CR' ||
trim(TO_CHAR(cur.crit_id)) ||
                  ') - ' || cur.MIN_SCORE || '),0) SC' ||
trim(TO_CHAR(cur.crit_id));
END IF;
IF cur.ID_TYPE_GR=3 THEN
      ttt1 := ttt1 || '*nvl((CR' || trim(TO_CHAR(cur.crit_id)) || ' -
' || cur.MIN_SCORE ||
                  '),0) SC' || trim(TO_CHAR(cur.crit_id));
END IF;
END LOOP;
      EXECUTE IMMEDIATE 'CREATE OR REPLACE FORCE VIEW K_Scores ' ||
ttt ||
      ') as SELECT SA_IDENT_NO, INS_KOD, RAI_KOD, SACM_SACHE ' ||
ttt1 ||
      ' FROM aud.K_ORG, (SELECT COUNT(SA_IDENT_NO) cnt FROM
aud.K_ORG)';

END;

/
```

Для выявления искомых закономерностей требуется использование нескольких методов и алгоритмов. В таком случае одни методы используются в начале моделирования, другие - на дальнейших этапах. Пример: для определения однотипных групп организаций применялся один из методов кластеризации, в результате организаций были разбиты на группы, каждой из которых, присвоен код; далее мы пользовались методом деревьев решений. Код группы (результат работы предыдущего метода) использовался для интерпретации полученных закономерностей. Выбор метода, на основе которого будет построена модель, должен осуществляться с учетом постановки задачи, особенностей набора исходных данных, специфики решаемой задачи, результатов.

4.3 **Исследование и методология расчета рисков.**

Прописываются коэффициенты для двух групп идентификаторов, описанных выше. Обозначим их как К1 и К2.

Также прописываются коэффициенты приоритета для конкретных идентификаторов.

Первая группа.

В первую группу, как было сказано быше, входят идентификаторы 1, 2 и 4.Для этой группы вычисление показателя риска производится следующим способом:

1. Допустим, коэффициенты присвоены следующим образом:

   Идентификатору 1 присвоен С1 коэффициент;

   Идентификатору 2 присвоен С2 коэффициент;

   Идентификатору 4 присвоен С4 коэффициент,

   где С1+С2+С4 = 1;

2. Вычисляются средние значения для идентификароторв 1, 2 и 4 в зависимости от сферы экономической деятельности. (Отметим эти значения как Р1, Р2 и Р4);

3. Вычисляются значения для идентификароторв 1, 2 и 4 в соответствии с конкретным налогоплательщиком (обозначим их как G1, G2 и G4);.

4. Вычисляется отклонение для конкретного налогоплательщика от соответствующего среднего значения по виду экономической деятельности с помощью следующей формулы :

   A1 = (G1-P1)/P1,

   A2 = (G2-P2)/P2,

   A4 = (G4-P4)/P4;

5. Вычисляется показатель риска для первой группы

   P1 = A1*C1 + A2*C2 + A4*C4;

Вторая группа.

В эту группу входят 3-й, 5-й, 6-й, 7-й, 8-й, 9-й, 10-й, 11-й и 12-й идентификаторы. для каждого идентификатора ползователь прописывает балл диапазона.

Например, для 3-го параметра

[0; 0.1] - 1 балл

[0.1; 0.2] - 5 балл

[0.2; 0.3] - 7 балл

и так далее.

Таким образом баллы прописываются и для 3-го,5-го,6-го,7-го,8-го,9-го,10-го,11-го и 12-го идентификаторов.

Определение показателя риска в этой группе производится следующим образом:

1. Вычисляются значения для 3-го,6-го,7-го,8-го,9-го и 11-го идентификаторов в зависимоцти от конкретного налогоплательщика;

2. 3-му и 10-му идентификаторам присваевается по 2 числа (болшьшее зачение соответствует болшему риску) в соответсвие с тем выполняют ли указанные условия;

3. Вычисляется показатель риска для каждого налогоплательщика в зависимости от прописанных баллов и коеффициэнтов. Например, показатель налогоплательщика в зависимости от идентификатора можно представить следующим образом:

идентификатор	3	5	6	7	8	9	10	11	12
показатель	0,25	2	50	15	0,5	125	1	-1,2	3

Таблица 5. Показатель риска

Баллы же и коэффициенты в зависимости от параметров выглядят так:

Идентиф.	3	5	6	7	8	9	10	11	12
диапазон	[0.2; 0.25]		[41;50]	[10;20]	[0.4;0.6]	[100;150]		[-1.1;-1.3]	[2;+]
балл	3	2	2	5	4	7	1	6	2
Коеффиц.	0.02	0.5	0.03	0.05	0.15	0.1	0.04	0.08	0.03

Таблица 6а. Показатель риска

Исходя из этого показатель риска для второи группы будет:

P2 = 3*0.02 + 2*0.5 + 2*0.03 + 5*0.05+ 4*0.15 + 7* 0.1 + 6*0.08 + 2*0.03 = 3.25.

Итоговое значение риска вычисляется по следующей формуле:

P= P1*K1 + P2*K2;

где K1 к K2 коеффициенты группы.

4.4 Проверка и оценка моделей

Проверка модели подразумевает проверку ее достоверности или адекватности. Эта проверка заключается в определении степени соответствия модели реальности. Адекватность модели проверяется путем тестирования. Адекватность модели - соответствие модели моделируемому объекту или процессу. Понятия достоверности и адекватности являются условными, поскольку мы не можем рассчитывать на полное соответствие модели реальному объекту, иначе это был бы сам объект, а не модель. Поэтому в процессе моделирования следует учитывать адекватность не модели вообще, а именно тех ее свойств, которые являются существенными с точки зрения проводимого исследования. В процессе проверки модели необходимо установить включение в модель всех существенных факторов. Сложность решения этой проблемы зависит от сложности решаемой задачи. Проверка модели также подразумевает определение той степени, в которой она действительно помогает менеджеру при принятии решений.

Оценка модели подразумевает проверку ее правильности. Оценка построенной модели осуществляется путем ее тестирования.

Тестирование модели заключается в "прогонке" построенной модели, заполненной данными, с целью определения ее характеристик, а также в-проверке ее работоспособности. Тестирование модели включает в себя проведение множества экспериментов. На вход модели могут подаваться выборки различного объема. С точки зрения статистики, точность модели увеличивается с увеличением количества исследуемых данных. Алгоритмы, являющиеся основой для построения моделей на сверхбольших базах данных, должны обладать свойством масштабирования.

Построенные модели рекомендуется тестировать на различных выборках для определения их обобщающих способностей. В ходе экспериментов можно варьировать объем выборки (количество записей), набор входных и выходных переменных, использовать выборки различной сложности.

Выявленные соотношения и закономерности должны быть проанализированы экспертом в предметной области - он поможет определить, как являются выясненные закономерности (возможно, слишком общими или узкими и специфическими).

Для оценки результатов полученных моделей следует использовать знания специалистов предметной области. Если результаты полученной модели эксперт считает неудовлетворительными, следует вернуться на один из предыдущих шагов процесса Data Mining, а именно: подготовка данных, построение модели, выбор модели. Если же результаты моделирования эксперт считает приемлемыми, ее можно применять для решения реальных задач.

Выбор модели

Если в результате моделирования нами было построено несколько различных моделей, то на основании их оценки мы можем осуществить выбор лучшей из них. В ходе проверки и оценки различных моделей на основании их характеристик, а также с учетом мнения экспертов, следует выбор наилучшей. Достаточно часто это оказывается непростой задачей.

Основные характеристики модели, которые определяют ее выбор, - это точность модели и эффективность работы алгоритма.

После тестирования, оценки и выбора модели следует этап применения модели. На этом этапе выбранная модель используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining.

Итак, рассмотрим принципы работы созданной нами программы.

Раздел 1 -" декларации на прибыль

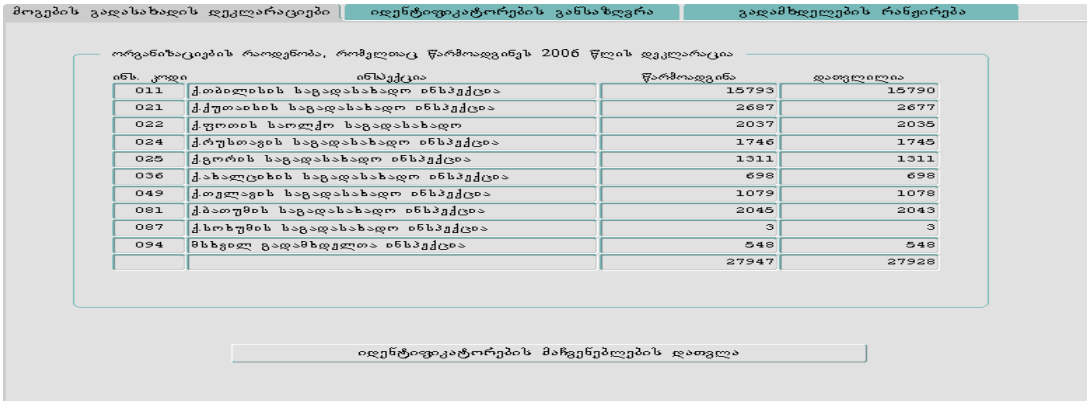

Рис.10 Информация о предъявленных декларациях

Данный раздел посвящен информации о налоге на прибыль, сгруппированной по районным налоговым инспекциям, и показывает сколько организаций предъявили декларации в 2007 году.

1 столбец - номер региональной налогой инспекции;

2 столбец - название регионалнои налогой инспекции;

в 3 столбце показана информация о количестве организаций , предъявивших данную декларацию;

в 4 столбце показана информация о количестве организаций, чьи данные подсчитаны на момент работы с программой.

Нажатием на кнопку "подсчитать показатели индетификаторов" получаем информацию о всех налогоплательщиках, предъявивших декларацию о налоге на прибыль за 2007 год.

Раздел 2 -" определиение идентификаторов"

При переходе на данную страницу откривается следующее окно:

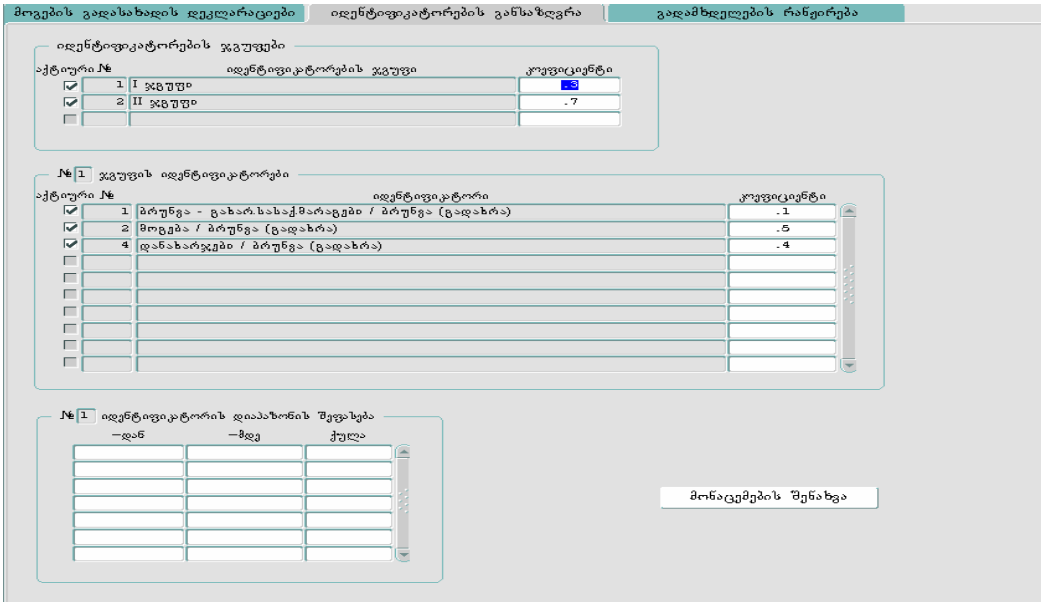

Рис.11 Группы идентификаторов

В разделе "группы идентификаторов" определяются идентификаторы для существующих двух групп.

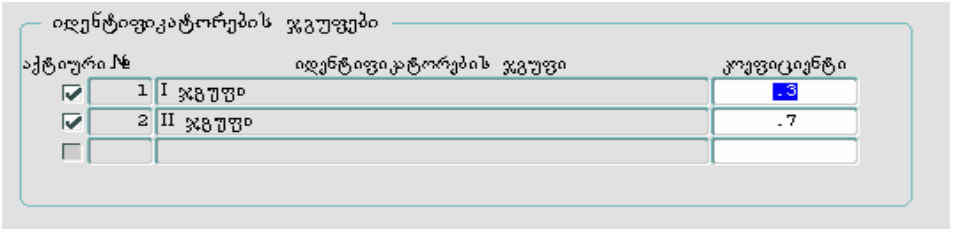

Рис.11а Группы идентификаторов

Вписывая номер группы в соответствующую ячейку, изменяется список идентификаторов во втором отлеле, так для первой группы выводятся идентификаторы , показанные на рисунке

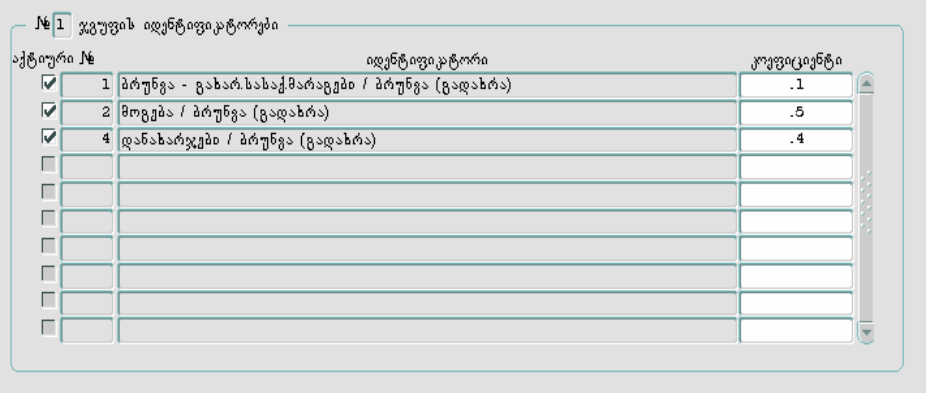

Рис.11b Группы идентификаторов

Для второй группы идентификаторов список будет представлен в следующем виде соответственно:

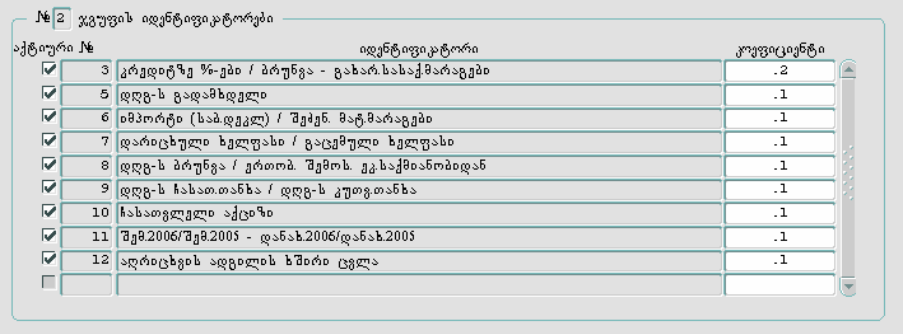

Рис.11c Группы идентификаторов

Изменяя значение идентификатора в ячейке "диапазон идентификатора", вносится диапазон для каждого идентификатора. Так , например, для третьего коеффициента таблица предтавляется в следующем виде:

Рис.11d Диапазон идентификаторов

Где дается возможность указать желаемый диапазон и присвоить соответствующий балл. Далее , нажатием на кнопку "сохранить данные" , происходит запись в базу данных.

Раздел 3 -" Ранжирование налогоплательщиков"

Здесь проиходит конечный подсчет суммы баллов для каждого налогоплательщика.

Раздел "фильтрация"

Программа дает возможность фильтровать выводимую информацию по идентификационному номеру, виду экономической деятельности, инспекции и по диапазону оборота.

Рис.11e Раздел фильтрации

В следующем разделе отображается иформация конкретно о среднем показателе экономическои деятельности, т.е. для 1,2 и 4 -го идентификаторов.

Например, при выборе вида экомоническои деятельности "строительство" получим следующюю информацию:

საშუალო მაჩვენებელი №1, 2 და 4 იდენტიფიკატორებისათვის

მშენებლობა

1 – .55863 2 – .08575 4 – .94397

Рис.11f раздел среднего показаделя для групп

После установки желаемых фильтров, нажатием кнопки "вывод информации", таблица заполняется. На экране появляется список налогоплательщиков, отсортиванный по полученным баллам в порядке уменьшения.

В распечатанном виде таблица выглядет так:

საიდენტ.კოდი	ორგანიზაციის დასახელება	ინს.	რ-ნი	საქმიანობის სახე	ქულა	1
205170955	მაქს ფლექსი	011	006	კერძო საოჯახო მეურნეობა დაქირავებული	1568.99	1
231258359	შპს "ბესიონი-პლანტს"	049	049	აგრარია, ნადირობა, მეტყევეობა და თევზჭე	1405.53	6.73
204480560	შპს პრიმა-დიდომი	011	002	ვაჭრობა, საყოფაცხოვრებო მომსახურება დ	1401.49	.86
212677922	სს ხის დამამუშავებელი კომბინატი	021	021	გადამამუშავებელი მრეწველობა	1401.21	.58
204988638	"ლატანი"	011	006	გადამამუშავებელი მრეწველობა	1401.13	.91
205023927	საზოგადოება "ლევან გეჩეჩილაძე" ფოფ ჯგუფიდი	011	006	ვაჭრობა, საყოფაცხოვრებო მომსახურება დ	1400.95	1
202252766	ჯი ემ პოლიპაკი	011	002	გადამამუშავებელი მრეწველობა	1400.9	.16
208140171	ფორა (ფოფ. ავტო 95)	011	007	ვაჭრობა, საყოფაცხოვრებო მომსახურება დ	1400.9	.34
202062123	"ბექი კომპანი"	094	094	გადამამუშავებელი მრეწველობა	1400.89	.23
204488660	იტალმობილი	011	005	ვაჭრობა, საყოფაცხოვრებო მომსახურება დ	1400.83	.26
201954288	ელექტროტექნიკური ქარხანა	011	002	გადამამუშავებელი მრეწველობა	1400.83	1
204935099	შპს კორპორაცია ქართული ღვინო	011	006	გადამამუშავებელი მრეწველობა	1400.82	.38
238742269	შპს მეინგარი	021	064	გადამამუშავებელი მრეწველობა	1400.8	.41
206079768	შპს აგა	011	007	კერძო საოჯახო მეურნეობა დაქირავებული	1400.8	.28
202236794	შპს ქუთაისის ავტომექანიკური ქარხანა (ფოფ შ	021	021	კერძო საოჯახო მეურნეობა დაქირავებული	1400.78	.12
205146625	შპს პრიზა ყოფილი ჯბბ	094	094	მშენებლობა	1400.77	.66
202200171	ჩინი პროდაქტ	011	002	კერძო საოჯახო მეურნეობა დაქირავებული	1400.74	.09
205056213	შპს გეო-ფილტრი	011	006	სხვა კომერციული საქმიანობა	1400.73	.22
204935972	"ტიტანიკი-2000"	011	006	სხვა კომერციული საქმიანობა	1400.72	.03
225392009	გარანტი-2006	024	039	მშენებლობა	1400.72	1.3

2	3	4	5	6	7	8	9	10	11	12
0	1501.81	6349.17	2	0	1.17	0	0	1	0	0
0	.83	19.04	2	2.28	1.43	0	1.53	1	0	0
0	.86	4.42	1		1.9	0	0	1	-38.01	0
0	.91	3.64	1		2.92	0	0	1	-1.91	0
0	.98	2.71	2		5.65	0	0	1	3.22	0
0	.93	1.71	2	0	.84	.04	4.42	1	18.77	0
0	1	1.71	2	1.11	1	1	3.14	1	-6.35	0
0	.84	1.72	2	5.21	.45	.93	1.82	1	137.22	0
0	.83	1.37	2	1.19	1.02	.97	2.27	1	-.17	0
0	.89	1.22	2	2.12	1	.99	3.03	1	550230.05	0
0	.9	2.05	2	0	0	1	.22	1	-.09	0
0	.91	1.31	2	3.09	.75	.99	2.22	1	-.08	0
0	.93	1.51	2		1.18	1	1.08	1	-.31	0
.01	.85	.98	2	2.38	1.2	.99	1.11	1	-.1	0
0	.99	1.12	2		1.18	.99	.91	1	-2.43	1
0	.99	2.73	2		.71	.76	14.12	1	0	0
0	.86	.99	2	1.33	1	1	1.07	1	0	0
0	.82	1.26	2	1.88	1	1	1.66	1	-.63	0
0	.87	1.1	2		0	1	8.24	1	0	0
0	.94	2.07	2	.58	0	0	0	1	0	0

Рис.11g Список ранжированных налогоплательщиков

Где в первом столбце показан идентификационный номер налогоплательщика, во втором название организации, в третьем код инспекции, в четвертом код подчиненнои инспекции, в пятом бид экономическои деятельности. Далее следуют показатели для каждого идентификатора в отдельности, пронумерованные в соответствии с номерами идентификаторов. Таким образом, полученная информация представляется следующим графиком.

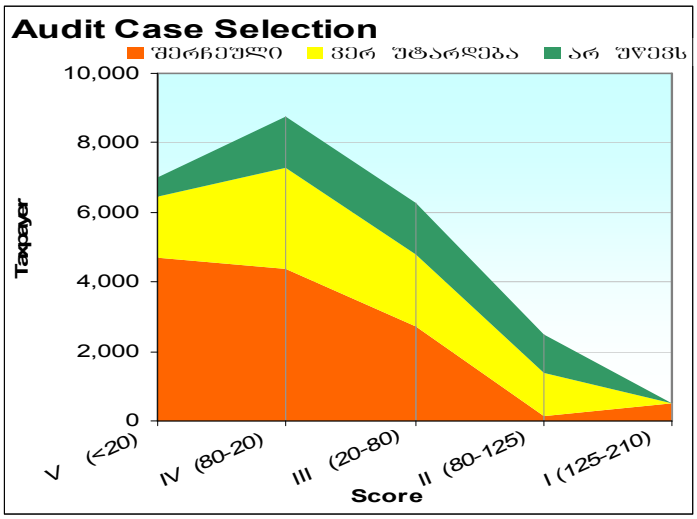

Рис.12 Графическое представление множества отобранных налогоплательщиков.

Анализируя представленные данные, мы видим, что налогоплательщики разбиты на 5 групп, в каждую из которых попали организации в соответствии с установленным диапазаном баллов.

Первая группа, в силу своей важности, полностью подлежит контрольной проверке. В остальных группах производится выборка, наиболее приоритетных с точки зрения контроля налогоплательщиков, в зависимости от имеющегося резерва свободных аудиторов в районных налоговых инспекциях.

	Подлежащие Конрольной Проверке	Проверка не проводится	Не подлежит проверке
V (<20)	4,708	1,750	543
IV (80-20)	4,375	2,917	1,458
III (20-80)	2,708	2,083	1,458
II (80-125)	125	1,250	1,125
I (125-210)	500	0	0

Таблица 6. Группы налогоплательщиков

Итак, как видно из следующего рисунка, ранжирование налогоплательщиков, т. е. Создание моделей происходит в центральном оффисе. С помощью

квалифицированных специалистов, путем изменения коэффициентов, с последующим пересчетом данных посредством хранимых проседур находится модель , наиболее точно отражающая действительность.

Таким образом, полученный список ранжированных налогоплательщиков поступает в районные налоговые инспекции, где производится распределие дел между аудиторами, принимается решение о проверке дел с низким приоритетом и происходит управление процессом проведения аудита. Заключительным звеном процесса является выезд аудитора на контрольную проверку, проведение аудита и подготовке формы результатов аудита.

Рис.13 Процессы аудита

Заключение

В заключение можно сказать , что государственное управление налоговыми отношениями осуществляется в рамках системы налогового администрирования, включающей в себя налоговое планирование, налоговое регулирование и налоговый контроль. При этом налоговый контроль составляет центральное звено в системе налогового администрирования, поскольку направлен на решение как фискальных, так и регулирующих задач, выступает в качестве ее наиболее активной, мобильной и результативной подсистемы.

В итоге была создана система компьютерных моделей, воспроизводящих условия и последствия применения различных вариантов налогообложения и имеющих в своей основе широкую информационную базу. На ней, как на испытательном стенде, были протестированы все возможные сценарии проектируемого налогового механизма.

Значимость диссертации заключается в разработке методологических подходов, механизмов и рекомендаций по созданию эффективной системы налогового контроля и организации налоговых проверок на основе автоматизации контрольных процедур и технологий.

Данная информационно-вычислительная система, базирующаяся на информации, извлекаемой (с помощью методологий Data Mining) из данных, поступающих в налоговые инспекции, помогает в условиях полной неопределенности, провести работу по поиску достаточной информации, с тем, чтобы получить возможность выявить зоны рисков налогоплательщиков.

Таким образом, проведенное исследование показало, что технологии Data Mining могут успешно применяться для выявления скрытых тенденций. При этом следует отметить, что в отличие от других методов поддержки принятия решений технологии Data Mining обладают гораздо более высокой степенью интеллектуальности и хорошей масштабируемостью, позволяя в значительной степени автоматизировать анализ данных.

Литература:

1. Налоговый кодекс Грузии.

2. საქართველოს ფინანსთა სამინისტროს ბრძანება №228 “გადამხდელთა პირადი აღრიცხვის ბარათების წარმოების წესის შესახებ”//2005 წ. 7 აპრილი, ქ. თბილისი

1. Л. Л. Винокуров, Д. В. Леонтьев, А. Ф. Гершельман. СУБД ADABAS - основа универсального сервера баз данных // *СУБД*. - 1997. - №2. - С. 36-40.
2. Н. Е. Емельянов. Введение в СУБД ИНЕС. - М.: Наука, 1988.
3. Н. Кречетов, П. Иванов. Продукты для интеллектуального анализа данных // *ComputerWeek-Москва*. - 1997. - N 14-15. - С. 32-39.
4. А. А. Сахаров. Концепции построения и реализации информационных систем, ориентированных на анализ данных // *СУБД*. - 1996.- N 4. - С. 55-70.
5. E. F. Codd, S. B. Codd, C. T. Salley. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. - E. F. Codd & Associates, 1993.
6. J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals // *Data Mining and Knowledge Discovery*. - 1997. - N 1. - P. 29-53.
7. D. Hackathorn. Reinventing Enterprise Systems Via Data Warehousing. - Washington, DC: The Data Warehousing Institute Annual Conference, 1995.
8. W. H. Inmon. Building The Data Warehouse (Second Edition). - NY, NY: John Wiley. - 1993.
9. K. Parsaye. New Realms of Analysis: Surveying Decision Support // *Database Programming & Design*. - 1996. - N 4. - P. 26-33.
10. N. Raden. Данные, Данные и только данные // *ComputerWeek-Москва*. - 1996. - №8. - С. 28.

11. Bhaskar Himatsinga, Juan Loaiza Oracle Corporation . All About Oracle Database Fragmentation.
12. Oracle Database Administrator's Guide. Partitioning in Data Warehouses, 2003

13. Tom Kyte ‘Expert one on one.’

14. Артемьев В. Что такое BUSINESS INTELLIGENCE? //Открытые системы, 2003, №4.

15. Архипенков С., Голубев Д., Максименко О. Хранилища Данных. От концепции до внедрения/ Под общ. Ред. Архипенкова С.Я. // М.: ДИАЛОГ-МИФИ, 2002 г.

16. Архипенков С. Хранилища данных. От концепции до внедрения. 2002 г.

17. Банерджи Д. Витрины данных и системы генерации отчетов //Tate Bramald Consultancy Co.

18. Бергер Ч. Data Mining от Oracle: настоящее и будущее/ Oracle's Data Mining Solutions – Oracle OpenWorld White Paper// Источник: Конференция OOW2000, San Francisco, доклад 134 /20 января 2001 г.

19. Большаков П.С. Уникальные возможности STATISTICA Data Miner

20. Галахов И. В., Волков И. Ю. Архитектура современной информационно-аналитической системы // Директор ИС №3, 2002г.

21. Гик Дж., ван. Прикладная общая теория систем. - М.: Мир, 1981

22. Inmon W. H. Building the Data Warehouse. New York: John Wiley & Sons, Inc.

23. E. F. Codd, S.B.Codd. Providing OLAP. On-line Analytical Processing to User-Analists: An IT Mandate. C. T. Salley, E. F. Codd & Associates, 1993

24 R. Kimball. The Data Warehouse Toolkit. Practical Techniques for Building Dimansional Data Warehouses.

25. Tom Kyte 'Expert one on one.'

26. Алексей Федоров, Наталия Елманова КомпьютерПресс 4'2001;

27. "Принципы проектирования и использования многомерных баз данных (на примере Oracle Express Server)" А.А.Сахаров, СУБД, №3, 1996;

28. 1. Bhaskar Himatsinga, Juan Loaiza Oracle Corporation . All About Oracle Database Fragmentation

29. Джиовинаццо В. Построение Хранилища данных для web-среды. Октябрь 2002 год

30. Джукич Н. Разработка одношагового Хранилища данных для поставщика услуг финансовых приложений. Январь 2002 г.

31. Дюк В., Самойленко А., Data mining: учебный курс. - СПб: Питер, 2001г.

32. Киселев М., Соломатин Е. Средства добычи знаний в бизнесе и финансах. - Открытые системы, № 4, 1997.

33. Кнут Д. Искусство программирования, том 1. Основные алгоритмы, 3-е изд.: - М.: "Вильямс", 2000г.

34. Комафорд К. Корпоративная отчетность: Серверная архитектура для распределенного доступа к информации. // Открытые системы, 1999, 2.

35 Кречетов Н., Иванов П. Продукты для интеллектуального анализа данных // ComputerWeek-Москва. - 1997. - № 14-15.

36. Львов В. Создание систем поддержки принятия решений на основе хранилищ данных, СУБД//1997, №3.

37. Островский Е. В. Порядок разработки ETL- процессов//7 октября 2004 .

38. Пржиялковский В. В. Сложный анализ данных большого объема: новые перспективы компьютеризации // СУБД. - 1996. - № 4.

39. Пройдаков Э. Что такое Data Mining?// PC Week/RE 99/26.

40. Раден Н. Данные, данные и только данные // ComputerWeek-Москва. - 1996.

41. Самохвалов Р. Системы поддержки принятия решений Oracle (Часть 1) // Oracle СНГ 2003г.

42. Сахаров А.А. Концепции построения и реализации информационных систем, ориентированных на анализ данных .

43. Сахаров А.А., Принципы проектирования и использования многомерных баз данных //СУБД №3, 1996

44. Семененко А.В. Витрины данных – новая технология обработки, анализа и хранения информации для систем поддержки принятия решений//(статья с сайта www.nit7.artdesign.ru).

45. Смирнов Н. Система для «Манчестера» 18.08.2003// Еженедельник "Computerworld", #30, 2003 год // Издательство "Открытые системы"

46. Смол Р., Two Crows Corporation// (источник: http://www.management-magazine.ru), Апрель, 2003 г.

47. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация. ..Том.1: Пер. с англ. М. Вильямс, 2001г.

48. Туманов В. Data Warehouse: с чего начать?// PC Week/RE 1998 г.

49. Туо Дж. Каждому пользователю - свое представление данных // ComputerWeek-Москва, 1996, № 38

50. Туо Дж. Инструменты для анализа информации на настольных ПК // ComputerWeek-Москва,1996, № 38.

51. Федоров А., Елманова Н. Введение в OLAP: часть1. Основы OLAP// КомпьютерПресс 4’ 2001

52. Чаудхури С., Дайа У., Ганти В. Технология баз данных в системах поддержки принятия решений //22 января 2002г,(статья с сайта www.citforum.ru).

53. Alalouf C. Hybrid OLAP// St. Laurent, Canada: Speedware Corporation Inc., 1997.

54. Almeida M. S., Ishikawa M., Reinschmidt J., Roeber T. Getting Started with Data Warehouse and Business Intelligence// IBM Red Books

55. An Introduction to Multidimensional Database Technology// Kenan Systems Corporation, 1995.

56. Boulding K. E. General Systems Theory. The Skeleton of Science//Management Science, 2, 1956.

57. Caserta J., Kimball R. The Data Warehouse ETL Toolkit//Practical Techniques for Extracting, Cleanin September 13, 2004.

58. Celko Trees in SQL// Intelligent Enterprise, October 20, 2000.

59. Codd E. F., Codd S.B. Providing OLAP. On-line Analytical Processing to User-Analists: An IT Mandate// C. T. Salley, E. F. Codd & Associates, 1993.

60. Data Warehouse Issues// Butler Group Co., UK 2004.

61. Demarest M. Building the Data Mart // DBMS. , July,1994.

61. Gartner Research, "BI Magic Quadrants: A 'Recession-Proof' Market Challenged", 17.07.2001

63. ORACLE. Data Mining. Administrator’s Guide//10g release2(10.2) B1433B-01, June, 2002.

64. Parsaye K. Surveying Decision Support: New Realms of Analysis // Database Programming and Design. - 1996. - № 4.

65. Pendse N. OLAP Architectures// The OLAP Report, (http://www.olapreport.com/Architectures.htm#top).

66. Oracle Database Administrator's Guide. Partitioning in Data Warehouses, 2003

67. Poly Analyst Решения для лучших в бизнесе// Компания Мегапьютер

68. Абулашвили Х., Маградзе Е.: Построение и реализация Хранилищ Данных,- “INTELECT”/ -2006| №1(24)

69. ხ. აბულაშვილი, ე. მაღრაძე: ოპერაციების კვლევა Data Mining ტექნოლოგიის საფუძველზე, -2006 | No.4(11) [2006.12.30], (http://gesj.internet-academy.org.ge/gesj_articles/1234.pdf)- - უკანასკნელად იქნა გადამოწმებული - 5.08.2008

70. Абулашвили Х., Маградзе Е.: Фрагментация таблиц в Базе Данных ORACLE,- “INTELECT”/ -2006| №1(25)

71. Маградзе Е., Абулашвили Х.: Оптимтзация запросов и материализованные представления в Базе Данных ORACLE,- “Georgian Engineering News” /-1'06

72. www.mof.ge- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

73. www.sql.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

74. www.olap.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

75. www.oracle.com- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

76. www.metalink.com- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

77. www.cittforum.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

78. www.ibm.com/bi- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

79. www.sas.com/datamining- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

80. www.spss.com- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

81. www.StatSoft.com- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

82. www.otr.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

83. www.sqlinfo.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

84. www.interface.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

85. www.olapreport.com- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

86. http://www.iso.ru/journal/articles/257.html- უკანასკნელად იქნა გადამოწმებული- 5.08.2008

87. http://www.hyperion.ru/content.phtml?section_id=97- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

88. http://olaplib.contourcomponents.ru/- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

89. www.iteam.ru- უკანასკნელად იქნა გადამოწმებული - 5.08.2008

90. www.interface.ru/datamining/datamining.htm- უკანასკნელად იქნა გადამოწმებული - 5.08.2008)

91. www.taxdep.ge - უკანასკნელად იქნა გადამოწმებული - 5.08.2008.