

Кыргызский Национальный Университет им. Ж. Баласагына

Институт Интеграции Международных образовательных программ

Кыргызско-Американский Факультет компьютерных технологий и Интернет

Курсовой проект

По курсу: «Информационные системы управления»

Тема: «Группировка пользователей ресурсов сети корпоративного доступа в Интернет с использованием методов корреляционного анализа»

Выполнил: Похилько А.Ф.

Группа КИС -4-00

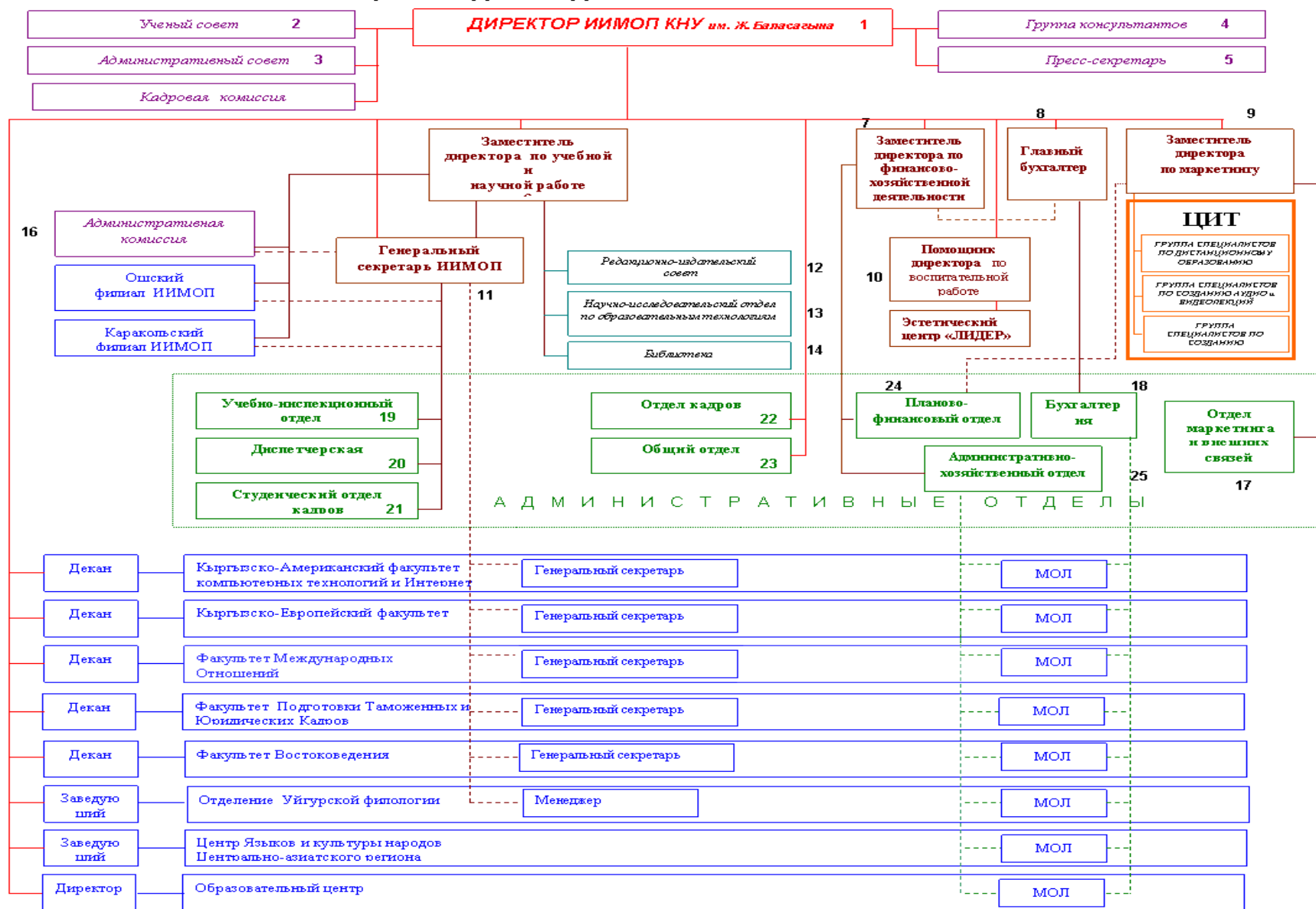
Приняла: Абдрахимова Н.Д.

Бишкек 2004г.

Содержание:

ВВЕДЕНИЕ	5
ТЕОРИЯ ИСУ	6
Жизненный цикл создания и развития ИСУ	6
Цели и задачи ИСУ	7
СХЕМА РАБОТЫ ИСУ	8
<i>Агрегатор</i>	9
<i>Фильтр</i>	9
<i>Анализатор</i>	9
<i>Интерпретатор</i>	9
ТЕОРИЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА	10
ПРЕДОБРАБОТКА И ОЧИСТКА ДАННЫХ	10
ИНТЕРПРЕТАЦИЯ ДАННЫХ	10
ПРОГРАММНАЯ РЕАЛИЗАЦИЯ.	12
Модули программы	12
<i>matrix.php</i>	12
<i>form.php</i>	12
<i>dbconnect.php</i>	12
<i>anz.php</i>	12
РЕЗУЛЬТАТЫ АНАЛИЗА.....	13
ВЫВОД	14
ЛИТЕРАТУРА.....	15

ИНСТИТУТ ИНТЕГРАЦИИ МЕЖДУНАРОДНЫХ ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ КНУ им. Ж. Баласагына



Введение

Любое решение, принимаемое человеком на чем-то основывается. Чаще всего эти основания предоставляют результаты анализа какого-либо объекта или области. По своей природе человек хорошо выявляет явные зависимости, и практически неспособен обнаруживать скрытые, неявные зависимости между фактами и событиями. В этой ситуации на помощь приходит корреляционный анализ и вычислительная техника. Вычислительная техника позволяет провести вычисления в кратчайшие сроки, без потери актуальности анализа.

Немаловажную роль играет корреляционный анализ, с помощью которого выявляются неявные зависимости. Корреляционный анализ возможен только в том случае, когда исследуемые показатели приведены к единой системе координат, в особенности это касается координаты "время". Для вычисления корреляционных зависимостей требуется строгое соблюдение единого масштаба времени. Выявление неявных зависимостей между происходящими процессами позволяет более точно определить причинно-следственные связи, способствует принятию взвешенных решений и помогает совершенствовать деятельность предприятия с целью повышения качества. [Сергей Текотев]

Теория ИСУ

Компьютерная информационная система состоит из людей, процедур, данных, программ компьютеров.

Таким образом, можно выделить следующие элементы данной информационной системы:

1. **Люди** – пользователи, которые будут работать с информационной системой.
2. **Компьютеры** – компьютеры, на которых будет установлена информационная системами, за которыми будут работать пользователи.
3. **Данные** – те данные, которые должна будет обрабатывать информационная система. В моем случае, это последовательность работ (их описание и продолжительность).
4. **Программа** – непосредственно сама программа, разработанная мною.
5. **Процедуры** – процедуры, использующиеся в программе для расчетов при определении критического пути.

Жизненный цикл создания и развития ИСУ

Под жизненным циклом системы обычно понимается непрерывный процесс, который начинается с момента принятия решения о необходимости создания системы и заканчивается в момент ее полного изъятия из эксплуатации.

Современные сети разрабатываются на основе стандартов, что позволяет обеспечить, во-первых, их высокую эффективность и, во-вторых, возможность их взаимодействия между собой.

Вообще говоря, все стандарты на информационные системы (как и на любые системы вообще) можно разбить на следующие два основных класса:

- Функциональные стандарты, определяющие порядок функционирования системы в интересах достижения цели, поставленной перед нею ее создателями.
- Стандарты жизненного цикла, определяющие то, как создается, развертывается, применяется и ликвидируется система.

Модели, определяемые стандартами этих двух классов, конечно же взаимосвязаны, однако решают совершенно разные задачи и характеризуются принципиально различными подходами к их построению.

Поясним это на примере. Наиболее полной функциональной моделью системы является сама система, однако "биография" самой системы ни в коем случае не может рассматриваться в качестве модели ее жизненного цикла. Куда ближе к модели жизненного цикла информационной системы является описание жизни живого существа, начиная с момента зачатия.



Рис1. Жизненный цикл ИСУ

Таким образом, жизненный цикл информационной системы охватывает все стадии и этапы ее создания, сопровождения и развития:

• предпроектный анализ (включая формирование функциональной и

информационной моделей объекта, для которого предназначена информационная система);

- проектирование системы (включая разработку технического задания, эскизного и технического проектов);
- разработку системы (в том числе программирование и тестирование прикладных программ на основании проектных спецификаций подсистем, выделенных на стадии проектирования);
- интеграцию и сборку системы, проведение ее испытаний;
- эксплуатацию системы и ее сопровождение;
- развитие системы.

Цели и задачи ИСУ.

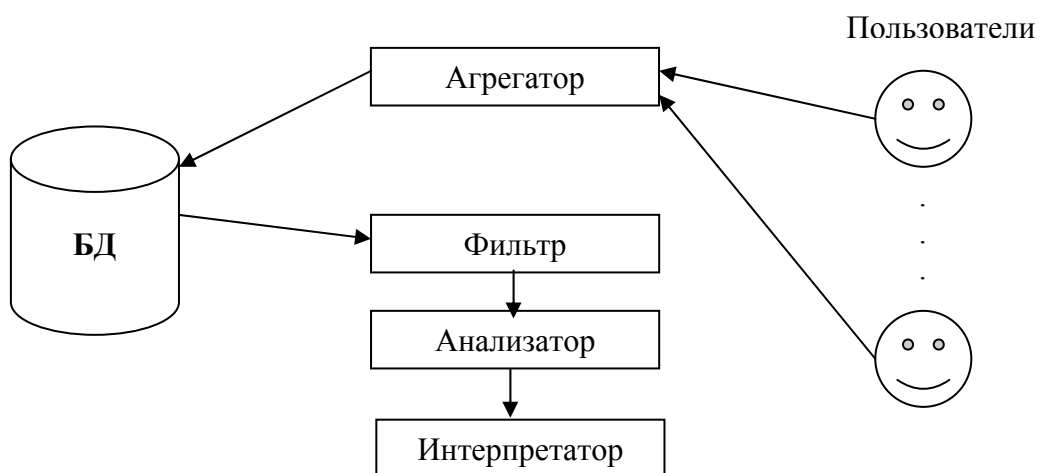
Основной целью DSS-системы является предоставление ЛПР информации, достаточной для принятия решения по управлению своей деятельностью. Наша аналитическая ИС позволяет выявить скрытые зависимости в пользовательской массе и принять соответствующие решения по управлению предоставлением Интернет-ресурсов (например, перераспределение объемов Интернет-трафика между группами пользователей).

Для достижения цели система решает следующие задачи:

1. Сбор и агрегация развернутых данных о пользовательской активности
2. Предоставление пользовательского интерфейса для выбора группы анализируемых данных
3. Нормализация и стандартизация данных для корреляционного анализа
4. Расчет неявных зависимостей путем корреляционного анализа
5. Вывод детального отчета с графиками

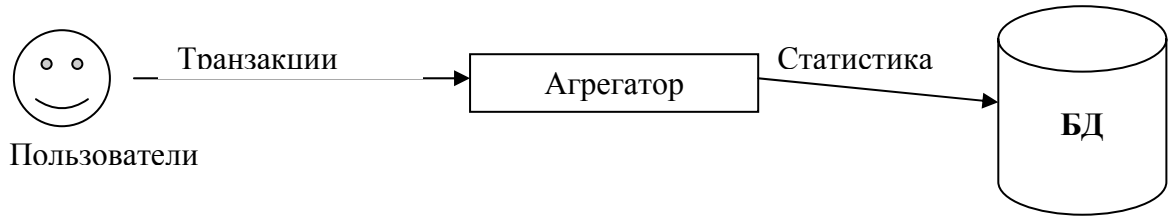
Схема работы ИСУ.

Представленная ИСУ имеет примерно схему работы:



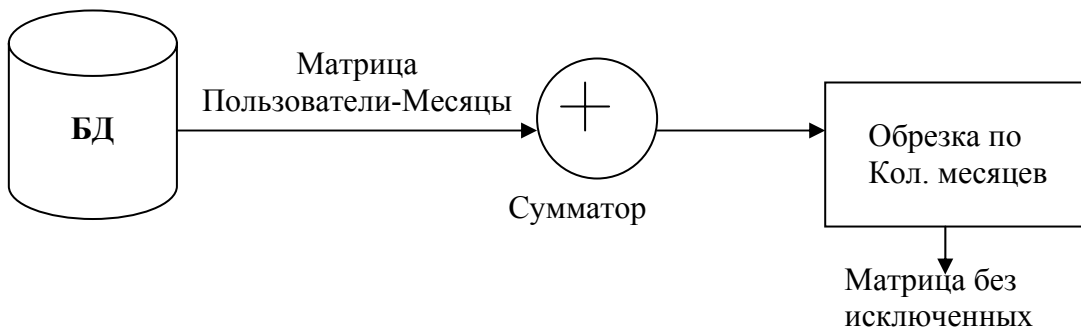
Агрегатор

Агрегатор получает на входе от системы корпоративного доступа в Интернет развернутые данные о запрошенной пользователями информации и ее объемах. На выходе агрегатора получают сводные данные об использовании пользователем трафика за каждый месяц.



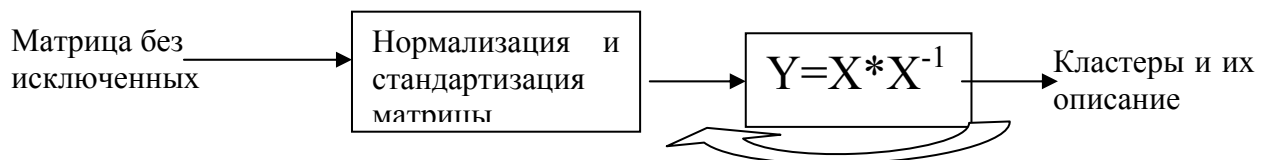
Фильтр

Фильтрация проводится для исключения из расчета пользователей, оказывающих незначительное влияние на общую статистику. Таковыми признаются пользователи, пользовавшиеся Интернет менее 8 месяцев в году.



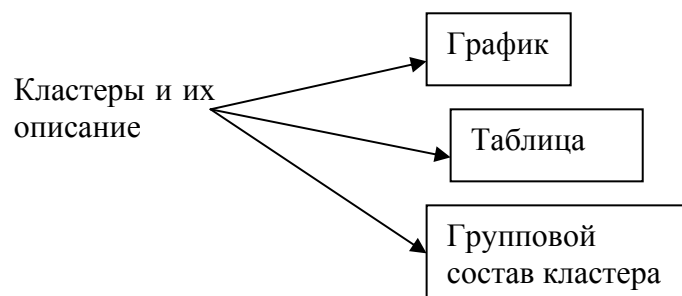
Анализатор

Анализатор производит операции корреляционного анализа. Итерации производятся до тех пор, пока количество кластеров не дойдет до указанного системе предела.



Интерпретатор

Интерпретатор превращает матрицу кластеров и массив их описаний в понятные человеку графики и таблицы с комментариями



Теория корреляционного анализа

Корреляционный анализ (Correlation analysis) - статистические методы обнаружения корреляционной зависимости между двумя или более случайными признаками или факторами. [encycl.yandex.ru]

Корреляционный анализ для двух случайных величин включает в себе:

1. построение корреляционного поля и составление корреляционной таблицы;
2. вычисление выборочных коэффициентов корреляции и корреляционных отношений;
3. проверка статистической гипотезы значимости связи.

Дальнейшее исследование заключается в установлении конкретного вида зависимости между величинами и составляет предмет задач, решаемых регрессионным анализом.

[www.basegroup.ru]

Предобработка и очистка данных

Для решения любой задачи корреляционного анализа задачи существует комплексная предобработка, в процессе которой осуществляется понижение размерности входных данных и/или устранение незначимых факторов. Очень часто в аналитических приложениях сосредотачивают усилия на механизмах анализа данных, не уделяя должного внимания задачам предобработки и очистки данных. Хотя именно плохое 'качество' исходных данных является одной из самых серьезных и распространенных проблем. Очевидно, что некорректные исходные данные приводят к некорректным выводам. А в связи с тем, что в большинстве случаев источником информации для аналитических систем является хранилище данных, в котором аккумулируются сведения из множества разнородных источников, острота проблемы существенно возрастает.

Необходимость предварительной обработки при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. При использовании же механизмов анализа, в основе которых лежат самообучающиеся алгоритмы, такие как нейронные сети, деревья решений и прочее, хорошее качество данных является ключевым требованием.

Очевидно, что исходные ('сырые') данные чаще всего нуждаются в очистке. В процессе этого восстанавливаются пропущенные данные, редактируются аномальные значения, вычитается шум, проводится сглаживание.

Но термин 'предобработка' можно трактовать шире, а именно как процесс предварительного экспресс анализа данных. Например, оценить фактор как значимый или нет, все ли факторы учтены для объяснения поведения результирующей величины и т.д. Для этих целей используются такие алгоритмы, как корреляционный анализ, факторный анализ, метод главных компонент, регрессионный анализ, спектральная обработка (сглаживание или вычитание шумов). [www.basegroup.ru]

В разработанной системе данные фильтруются по принципу значимости, так что пользователи с незначительным участием в формировании статистики не учитываются.

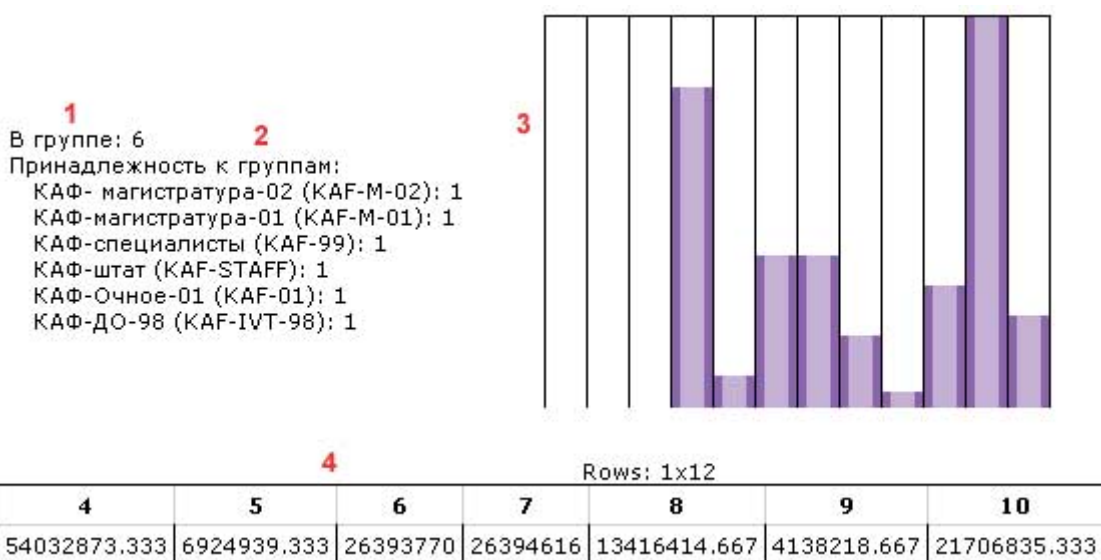
Интерпретация данных

Информация, найденная в процессе применения системы, должна быть нетривиальной и ранее неизвестной, например, скрытая зависимость между группой пользователей и характером использования ресурсов. Знания должны описывать новые связи между

свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность заключается в том, чтобы эти знания могли принести определенную выгоду при их применении. Знания должны быть в понятном для пользователя-нематематика виде. Например, проще всего воспринимаются человеком логические конструкции 'если ... то ...'. Более того, такие правила могут быть использованы в различных СУБД в качестве SQL-запросов. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду. [www.basegroup.ru]

В разработанной системе интерпретация производится на финальном этапе запроса к программе и выдает пользователю:

1. Количество пользователей, объединенных в группу
2. Распределение пользователей в логической группе по
3. График-гистограмму, визуально представляющий характеристику группы
4. Таблицу данных, по которой строился график



Проинтерпретированные данные анализатора

Программная реализация.

Система реализована с помощью программной технологии PHP, технологии СУБД MySQL. В качестве инструментального средства и среды разработки использовалась программа APCoder.

Программа разработана с использованием модульного программирования. Главный файл программы загружает модули и управляет вызовом функций этих модулей.

Программа предоставляет пользователю интерфейс выбора области данных для анализа и информирует его о текущем прогрессе и затратах времени на анализ.

Модули программы

matrix.php

Данный модуль был разработан по причине отсутствия в PHP встроенных функций обработки математических матриц. Модуль включает в себя функции:

- Findmax – поиск максимального элемента в строке матрицы
- join_row – объединение двух столбцов матрицы в один по правилам корреляционного анализа
- armul – умножение матрицы на матрицу
- rem_elem – превращение матрицы в нижнюю треугольную
- normalize – нормирование данных в матрице
- centralize – центрирование данных в матрице
- transp – транспонирование матрицы

form.php

Данный модуль содержит функции организации пользовательского интерфейса

- inout – выбор типа трафика : входящий/исходящий
- ttype – выбор зоны доступа
- years – выбор года, за который проводится анализ

dbconnect.php

Модуль организует соединение с СУБД по протоколу MySQL

anlz.php

Модуль содержит функции интерпретации результатов анализа в понятный для человека вид

- anlzGroup – выдача информации по принадлежности пользователей к группам
- analyze_row, graphRow – выдача графика-гистограммы и таблицы значений

Результаты анализа

Были проанализированы зависимости в период с апреля 2002 г по ноябрь 2003г по двум типам трафика (входящему и исходящему) и трем зонам доступа (локальная, Кыргызстан, внешний Интернет)

Результаты анализа оказались немного неожиданными, так как ожидалось нахождение зависимостей между использованием Интернет-трафика и группой, в которой находится студент. Однако, результат показал, что использование трафика ведется всеми пользователями вне зависимости от года обучения, принадлежности к факультету или группе. Тем не менее, выяснилось, что в любой год и в любой группе, как правило, пик использования ресурсов приходится на месяцы апрель или ноябрь, реже февраль. Также, выявляется четкая группировка пользователей *КАФ-специалисты* по активности лишь в определенные месяцы года.

Вывод

Бурное развитие средств вычислительной техники, информационных и телекоммуникационных технологий увеличили роль и значение информационных ресурсов и информационных систем в современном обществе, создало благоприятные предпосылки для широкого внедрения ИС во все сферы жизнедеятельности общества.

В результате работы была создана система, отвечающая поставленным требованиям и выдающая требуемый результат в понятном и удобном виде. Результаты анализа были записаны и предоставлены системному администратору КАФ-Интернет. Сама система также передана в отдел ИКТ ИИМОП КНУ для последующего использования в нуждах организации.

В результате выполнения курсовой работы автор значительно углубил знания, полученные на 2 курсе КАФ-Интернет по предмету «Линейная алгебра», а также изучил передовые технологии анализа больших объемов фактических данных.

Литература

1. Курс лекций по дисциплине «Информационные системы в Управлении».
2. Курс лекций по дисциплине «Бизнес-Математика».
3. Автоматизированный учебный курс <http://isu.online.kg>
4. Автоматизированный учебный курс <http://la.online.kg>
5. Кочергин В.П., Курс лекций «Линейная алгебра»