***ФГБОУ ВО «Кубанский государственный аграрный университет имени И.Т. Трубилина»,***

***Российская Федерация***

## **73 Интеллектуальные информационные системы и технологии. Лекция 2. Расчет моделей. Преобразование данных в информацию, а затем в знания. 2020-09-11**

**Заголовок:** Преобразование данных в знания: Расчет моделей и системно-когнитивный анализ в системе Эйдос

**Резюме текста:**

Лекция посвящена дисциплине "Интеллектуальные информационные системы и технологии", фокусируясь на расчете моделей и преобразовании данных в информацию и знания с математической точки зрения.

**I. Основные понятия и процесс трансформации данных:**

* **Данные:** Определяются как информация, рассматриваемая безотносительно к смысловому содержанию.
* **Информация:** Осмысленные данные. Смысл – знание причинно-следственных зависимостей.
* **Знания:** Информация, полезная для достижения цели. Требует постановки цели и нахождения способа использования информации о причинно-следственных связях.
* **Процесс:** Данные преобразуются в информацию, а информация – в знания.

**II. Расчет моделей и анализ данных:**

* **Подготовка:** Необходимо разработать справочники факторов, их значений и будущих состояний объекта моделирования, а также классификационные и описательные шкалы градаций. Исходные данные кодируются с их использованием, формируя базу событий.
* **Матрица абсолютных частот:** Первый этап анализа базы событий – расчет матрицы, где колонки – классы (градации классификационных шкал), а строки – признаки (градации описательных шкал). Значение ячейки – количество наблюдений объекта данного класса с данным признаком.
* **Ограничения абсолютных частот:** Не учитывают разное количество объектов в классах, что искажает сравнение.
* **Матрица относительных частот (вероятностей):** Для корректного сравнения абсолютные частоты преобразуются в относительные (делением на общее число признаков или объектов в классе). Вероятность – предел относительной частоты при бесконечном числе испытаний (абстракция), на практике работают с относительными частотами.
* **Количество информации (Мера Харкевича):** Логарифм отношения условной вероятности признака в классе к его безусловной (средней по выборке) вероятности. Позволяет оценить, насколько признак характерен для класса. Нулевое значение означает отсутствие информации.
* **Интегральные критерии:** Для оценки принадлежности объекта к классу суммируется количество информации по всем его признакам (аддитивный критерий). Используется для задач идентификации, классификации, прогнозирования.
* **Сравнение мер:** Рассматриваются и другие меры (Хи-квадрат, коэффициент Рои, коэффициент взаимосвязи), показывается их связь с мерой Харкевича и скалярным произведением векторов.

**III. Система Эйдос:**

* **Особенности:** Разработана автором, универсальна, не зависит от предметной области, находится в полном открытом доступе с исходными текстами, поддерживает многоязычный интерфейс (51 язык), устойчива к шуму в данных, использует графический процессор для ускорения расчетов (до 4000 раз), имеет нулевой порог входа.
* **История и применение:** Разрабатывается более 40 лет, применялась в различных областях (экономика, медицина, техника, психология, биология, АПК), защищено множество диссертаций.
* **Этапы работы:**
  1. Когнитивная (целевая) структуризация предметной области (неавтоматизированный этап): определение факторов (причин) и результатов их действия.
  2. Формализация: разработка шкал, кодирование данных, формирование базы событий.
  3. Синтез и верификация моделей: расчет матриц абсолютных, относительных частот и количества информации (7 типов моделей).
  4. Решение задач: идентификация, классификация, поддержка принятия решений, прогнозирование, исследование модели.

**IV. Пример и заключение:**

* Приводится пример с распознаванием студента/студентки по признакам (длинные волосы, брюки, мобильный телефон) для иллюстрации понятий количества информации и характерности признаков.
* Подчеркивается важность перехода от абсолютных частот к относительным и количеству информации для содержательного анализа.

**Детальная расшифровка текста:**

**I. Введение и основные понятия**

Здравствуйте. Здравствуйте. Сегодня у нас занятие по дисциплине "Интеллектуальные информационные системы и технологии". И мы рассматриваем вопрос: расчет моделей. Осуществляется расчет моделей конкретно с точки зрения математики. Рассматриваем, как данные преобразуются в информацию, а она в знание.

С чего мы начнем? С того, что чуть-чуть я вам напомню, что:

* **Данные** сложно определить, но можно определить их как информацию, рассматриваемую безотносительно к смысловому содержанию.
* **Информация** – это осмысленные данные.
* **Смысл** – это знание причинно-следственных зависимостей.
* А **знание** – это информация, полезная для достижения цели. То есть мы должны поставить цель и найти способ использовать эту информацию о причинно-следственных связях, которые у нас есть между прошлыми факторами и результатами их влияния на будущее поведение объекта моделирования. Если мы это можем сделать, то тогда этим самым информация преобразуется в знание.

**II. Процесс преобразования данных и расчет моделей**

**A. Подготовка данных и справочники**

Таким образом, для того, чтобы преобразовать данные в информацию, а её в знание, нужно сначала найти, разработать справочники факторов и их значений, справочники будущих состояний объекта моделирования. Ну я использую обычно универсальную терминологию, которая не динамичная и не статичная, а универсальная интерпретация. Просто говорю так: необходимо разработать классификационные и описательные шкалы и градации, закодировать исходные данные с их использованием, получить базу событий.

**B. Матрица абсолютных частот**

И потом в этих, в этой базе событий проанализировать её и найти там причинно-следственные связи между этими событиями. Таким образом, все вот эти базы данных: база исходных данных, база справочников прошлых и будущих событий, классификационных и описательных шкал градаций, база событий (закодированные исходные данные с помощью этих вот классификационных и описательных шкал градаций) и база, содержащая причинно-следственные взаимосвязи между событиями – всё вместе это информационные базы. Вот сейчас мы как раз и рассматриваем, как их получить.

Мы уже будем считать, что мы уже разработали справочники и базу событий. И теперь, первым делом, что мы делаем для того, чтобы исследовать причинно-следственные связи между этими прошлыми и будущими событиями? Первым делом, что мы делаем, это мы рассчитываем матрицу абсолютных частот. Сейчас я вам сделаю, чтобы было видно экран. Вы мне всегда подтверждаете, видно, не видно, потому что тут не всегда понятно. Вот. В Тимсе, в Тимсе там видно, рамочка появляется. Не совсем понятно. Видно, видно. Ну хорошо.

Вот. Рассчитывается, ребята, матрица абсолютных частот. Как она рассчитывается?

* Колонки этой матрицы представляют собой классы. То есть это обобщённые группы объектов наблюдения, соответствующие текущим или будущим состояниям объекта моделирования. Классы представляют собой градации классификационных шкал. Градации классификационных шкал. Их может быть много этих классификационных шкал. И их градации вот здесь расположены, они нумеруются подряд, кодируются от первого до последнего.
* А строчками являются градации описательных шкал, которые интерпретируются либо как значение свойств объекта моделирования, вот, либо как значение факторов. Я их называю признаками для простоты. Ну, градации описательных шкал – это признаки.

И вот, если у нас в каком-то объекте наблюдения, в каком-то наблюдении наблюдался объект, который объект обучающей выборки, который относился к житому классу, и у него был итый признак, то к соответствующему элементу матрицы абсолютных частот суммируется единичка. Вот так она и рассчитывается. Потом рассчитываются суммарные значения по строке, суммарные значения по колонке.

По колонке здесь надо уточнить. Есть два способа расчёта суммарных значений.

1. Такой более теоретически обоснованный, скажем так, это суммарное количество признаков. Вот сколько всего признаков было обнаружено в этой колоночке. Просто сумма этих n и житое элементов строки и столбца.
2. А можно рассматривать также суммарное количество объектов обучающей выборки, которые относятся к данному классу. Я почему их и выделил жёлтым эти колоночки, чтобы вот так вам показывать.

И рассчитывается также суммарное количество объектов обучающей выборки и суммарное количество признаков у этих объектов обучающей выборки.

**C. Ограничения абсолютных частот и переход к относительным**

В принципе, расчёт такой матрицы осуществляется практически во всех системах статистических, таких как SPSS или Статистика. Но на основе этой матрицы у них и проводятся дальнейшие расчёты, ну, скажем, кластерный анализ осуществляется и другие процедуры анализа.

В системе Эйдос же и в системно-когнитивном анализе, я считаю, что это не совсем правильно, так как это делается обычно. Я считаю, что необходимо перейти к другой матрице или преобразовать абсолютные частоты в относительные частоты, а потом преобразовать относительные частоты в количество информации. И потом уже вот эту матрицу, я их называю системно-когнитивные модели, в которых отражено уже количество информации (сейчас я к ним, к этой матрице подойду и подробнее объясню), уже её использовать для решения всех задач.

Почему нас не устраивает матрица абсолютных частот? По одной очень простой причине: объектов наблюдения по каждым классам, по разным классам, вообще говоря, разное количество. По одному классу одно количество объектов было обучающей выборки, объектов наблюдения, по другим классам – другое. Вот. Поэтому, если мы в одной колоночке видим три, а в другой пять, то это не означает, что в той колоночке, где пять, этот признак итый, скажем, итый признак, здесь мы видим три, а здесь пять, что этот признак более характерен для той колоночки, объектов той колоночки, где пять. Почему? А потому что там, где пять, у нас может быть 1000 наблюдений вот здесь, а там, где три, у нас может быть три наблюдения. То есть здесь у нас 100% объектов обладают этим признаком, а здесь там, скажем, там десятые доли процента объектов обладают этим признаком. Ну я так немножко, конечно, утрирую, такие большие различия редко наблюдаются, но, в общем, в частотах. Ясно, что сами по себе эти числа сравнивать нельзя, а надо перейти к относительным выражениям, к вероятностям или к относительным частотам.

**D. Расчет относительных частот и количества информации**

Делается это несложно. Для того, чтобы посчитать вероятность встречи итого признака у объектов житого класса, достаточно разделить количество встреч этого итого житого признака на число встреч, на число вообще признаков у объектов этого класса или на число объектов этого класса обучающей выборки. Получается, что у нас вот эта матрица условных и безусловных процентных распределений, она рассчитывается в двух вариантах. В одном варианте вот здесь в знаменателе находятся суммарное число признаков, а в другом – суммарное число объектов обучающей выборки по этому классу. Так оно и делается в системе Эйдос.

И сейчас я вам покажу дальше эти формулы для расчётов. Теперь, здесь мы уже получаем такие величины, которые не зависят от числа объектов обучающей выборки по классу. Относительные величины. Я называется здесь вот у нас в пособии, что это матрица условных и безусловных процентных распределений. Так я и называю эту матрицу обычно. Вот. А когда я про начал про неё рассказывать, я вам сказал, что тут надо рассчитать вероятности, перейти к относительным величинам, рассчитать вероятности. А сейчас показываю, что мы рассчитываем не вероятности, а относительные частоты. Соответственно, возникает вопрос о том, каким образом взаимосвязаны эти понятия: относительные частоты и вероятности.

Значит, у вас была ведь статистика, да, ребята? Была? Вы ведь можете голосом отвечать, мне голосом удобнее.  
(Ответ: Была)  
Вот. И вам на статистике, наверное, рассказывали, что вероятность – это есть предел, к которому стремится относительная частота при неограниченном увеличении числа испытаний. Было дело, да? А вот насчёт неограниченного увеличения, как вы думаете, что это значит? Ну это означает переход к бесконечности, это предельное понятие, связанное с предельным переходом, с большой предельной теоремой. В общем, короче говоря, я хочу, почему сейчас про это говорю? Потому что хочу, чтобы вы хорошо понимали, что это понятие абстрактное. Реально в жизни никогда неограниченное число наблюдений не наблюдается. Извините за каламбурчик. Ну то есть понятно, что число наблюдений всегда ограничено, причём, я бы сказал так, чаще всего довольно сильно ограничено.

И вот получается что? Что мы с вероятностями дела никогда не имеем. А фактически мы всегда имеем дело с относительными частотами. А вероятность представляет собой предел, к которому стремится относительная частота вот эта при увеличении числа событий до бесконечности, числа наблюдений до бесконечности. Такого не происходит, естественно. Ну чем больше наблюдений, тем меньше различия между относительной частотой и вероятностью. Это различие называют погрешностью иногда. Так вот эта погрешность в относительных величинах, если посчитать, ну я вам скажу, что можно сделать несложную программку, которая рассчитывает равномерное распределение случайной величины с равномерным распределением. И можно посчитать, как изменяется вероятность, э-э, как изменяется относительная частота, которую выбрасывает счётчик случайных событий, скажем, Random, Randomize там в Паскале. И как она стремится к некоторому пределу. Если, скажем, мы кидаем игральный кубик, имитируем в программе выбрасывание этого игрального кубика, то шесть граней есть. То там относительная частота выбрасывания, когда мы только одно испытание проведём, возьмём кубик этот, бросим, выпала тройка, например. То относительная частота выпадения тройки равна единице. Почему? Потому что один раз она выпала, один раз мы бросили. 1 / 1 будет 1. Второй раз мы бросили, выпала, допустим, двойка. Тогда относительная частота выпадения тройки становится равна 0,5 и двойки тоже 0,5. И вот так вот, если мы будем бросать и рассчитывать относительные частоты, то при большом числе испытаний мы будем видеть, наблюдать, что вот эта относительная частота стремится к 1/6. То есть к теоретической вероятности наблюдения этой вот выпадения любой из граней, когда выпадание их равновероятно.

Вот, собственно говоря, то, что я хотел сказать, что здесь, конечно, относительные частоты, но их отличие от вероятности при больших выборках, оно, в общем-то, несущественно. Но надо про него знать. И если мы говорим про вероятности, то нужно понимать, что вероятность – это абстрактные величины. Ну, наподобие, я вам скажу, наподобие бесконечно малых, наподобие геометрической точки, материальной точки. То есть большинство понятий геометрии и физики, они являются абстракциями, которые реально не наблюдаются. Когда мы берём интегралы, когда мы рисуем какие-то уравнения прямых там или кривых линий, это всё абстрактные математические объекты, которым в реальности ничего, строго говоря, не соответствует, но в какой-то, с какой-то погрешностью соответствует.

Вот. Так вот, дальше мы, я уже сказал, что двумя способами рассчитываем эту матрицу. И вот если мы здесь встречаем разные вероятности встречи признаков какого-то признака в разных классах: в первом, там, житом и W, то тут уже это существенная информация. Ну я всегда, когда студентам про это рассказываю, всегда привожу такой пример, очень наглядный. Вот, допустим, мы играем в такую игру. Очное у нас обучение, контактное. И за дверью стоит неизвестно кто: студент или студентка. А я вас спрашиваю: давайте задавайте вопросы, вернее, даю вам задание, задавайте вопросы про того, кто там стоит, на которые можно ответить да или нет. И будем догадываться, студент там или студентка. Ну, только касающиеся, конечно, имиджа, внешнего вида. И вот вы спрашиваете, сразу же первый же вопрос, который у вас возникает: волосы длинные? Я говорю: "Да, длинные". Вы получаете большой объём информации о том, что это похоже, что студентка. Почему? А вот почему, слушайте внимательно. Потому что вероятность встретить признак "длинные волосы" (итый вот этот признак – это признак "длинные волосы") в житом классе "студентки" равна почти 100%. Ну, допустим, 99% студенток имеют длинные волосы. Ну это близко к действительности. А вот в группе "студенты", будем считать, что это первая группа, длинные волосы имеет 1% студентов. И будем считать так условно, что у нас студентов и студенток одинаковое количество. Тогда получается так, что вероятность встретить длинные волосы у студенток 99%, вероятность встретить длинные волосы у студентов 1%. А суммарная вероятность встретить, то есть общая вероятность, безусловная вероятность по всей выборке, если их одинаковое количество, будет 98%, то есть около 50% получится. Потому что студенток примерно 50%, и они почти все с длинными волосами, получается, вероятность встретить признак "длинные волосы" по всей выборке около 50%.

И вот здесь мы смотрим: здесь вот 100, ну там 99%, здесь 1%, а в среднем 50. То есть это что значит? Что вероятность встретить этот признак в группе девушки в два раза выше, чем в среднем. У студентов крайне мала вообще, там сотые доли, да? Ну, в общем, вы поняли, крайне мала эта вероятность. У девушек она в два раза выше, чем в среднем, а у студентов она крайне мала. Получается что? Что мы, сравнивая вот эти вот вероятности встречи какого-то признака в одном классе и в другом, у объектов одного класса и другого, уже вполне обоснованно можем делать выводы. То есть это уже вполне корректно и возможно. Причём в качестве базы сравнения мы выбираем среднее по всей выборке для этого признака, вероятность его встречи по всей выборке. Ну да, почему это необходимо? Потому что мы здесь, допустим, видим 50%, но мы не знаем, много это или мало. Чтобы это понять, нам нужно с чем-то сравнивать. С чем мы можем сравнивать? Либо с другими процентами, процентами в других колонках, либо с тем, как это выглядит в среднем по всей выборке.

Значит, почему я говорю про процентное распределение? Значит, на самом деле здесь, конечно, относительные частоты, они превращаются в проценты, если умножить на 100. Вот. Ну это я люблю умножать на 100, потому что это и более привычно, это первая причина. А вторая причина, потому что лучше используется разрядная сетка. Можно одни целые оставить и сравнивать только целые, и уже есть смысл во всём этом, уже виден смысл. Вот, о чём несёт информацию тот или иной признак. Но самой информации здесь ещё нет, но мы уже можем её посчитать, сравнив эти вот относительные частоты в разных колонках. То есть информация здесь уже есть, но она в таком неявном виде. Она получается после сравнения вот этих вот условных и безусловных процентных распределений.

**E. Сравнение мер и интегральные критерии**

И вот само это сравнение, оно представляет собой работу, ребята. Вот представьте себе, что эта матрица, она небольшая, ну там, скажем, 10 колонок, 10 строчек. Тогда мы так визуально можем всё её охватить и проанализировать, скажем так, с помощью своего естественного интеллекта, и нам это не проблема. То есть мы видим смысл сразу этих вот всех чисел. Вот, понимаем их смысл. А вот если матрица побольше по размеру, ну, допустим, там 300 колонок и 1000 строк, то тогда уже как-то не очень хочется всю её просматривать и всё там анализировать. Вот. Но дело даже не только в этом. Ну то есть вручную это становится проблематичным. То есть это уже довольно большой труд будет. А дело в том, что мы-то не просто хотим это проанализировать, а мы хотим это использовать результаты этого анализа для решения различных задач. То есть тоже использовать его программно. Поэтому получается, что необходимо, из этих соображений, вот из этих предпосылок вытекает такой вывод, что необходимо эту матрицу условных и безусловных процентных распределений проанализировать автоматизированно. То есть эти все вот относительные частоты надо сравнивать их автоматизированно, а не вручную.

Как это сделать? Значит, я могу вам сказать, что если мы хотим сравнить два числа, вот, то есть, в общем-то, два честных способа и то есть один, один честный способ, да, и 1000 нечестных. 1000 нечестных и только один честный. Вот. Ну и сразу все спрашивают: "А какой честный?" Ну я говорю: "Ну так я и думал, что вы не знаете, какой честный". Это президент Рейган на дебатах выдал такую интересную вот пассаж. Так вот, ребята, всего есть два способа таких базовых, что ли, или простых, из которых можно разработать, на основе которых можно разработать неограниченное число способов таких, ну, скажем так, не очень разумных. Ну, например, я вам скажу, как вот можно получить число четыре? Можно 2 + 2, можно 2 \* 2, можно 6 - 2, можно 8 - 4, да? А можно 14 - 10. Так вот, когда я спрашиваю, каким способом можно получить число четыре, сразу все говорят: либо 2 \* 2, либо 2 в степени 2, 2². А то, что можно 1004 отнять 1000 и будет тоже четыре, никто не говорит. А почему? А потому что понятно, что таких способов их неограниченное количество, вот, и они не представляют собой особого интереса, вот, понимаете? То есть сразу возникает вопрос: а почему не 1005 - 1001 там, ну и так далее. Вот. Поэтому о них и не говорят.

Так же точно и здесь. То есть мы можем сравнить два числа либо вычев одно из другого и проанализировав результат больше нуля или меньше нуля, либо разделив одно на другое и определив отношение больше или меньше единицы или равно единице. Вот. Если мы, допустим, берём отношение, то когда вероятность встречи признака в группе равна вероятности встречи его по всей выборке, вот этот вариант, вот вы видите в табличке, третья табличка, приведены разные способы сравнения условных и безусловных процентных распределений, которые используются в системе Эйдос и соответствуют различным способам расчёта матриц, в которых отражены уже причинно-следственные связи. И вы уже, я вам несколько раз напоминал об этом и рассказывал подробно на первых, на первой паре, что расчётом этих матриц, содержащих причинно-следственные связи, силы и направления, заканчивается преобразование данных в информацию. То есть сейчас мы прямо конкретно рассматриваем математическую модель, как это можно сделать.

И мы видим, что можно, взяв отношение условной вероятности... Сейчас я вам пошире покажу, побольше. Вот. Так лучше видно с индексами, да, ребят?  
(Ответ: Да)  
Вот. Значит, берём логарифм этого отношения условной вероятности встречи признака итого признака в житой группе к безусловной вероятности встречи этого признака по всей выборке. И берём логарифм. А для чего мы берём логарифм? Потому что если у нас условная вероятность равна безусловной, то наличие этого признака ничего нам не говорит о принадлежности объекта с этим признаком к житой группе. Потому что вероятность его встретить... Ну я могу привести пример со студентом и студенткой. Ну, например, я вам говорю: "А у неё есть мобильный телефон". Вы говорите: "Ну и что?" Так мы интуитивно понимаем, что он есть у всех: и в группе студентки он есть у всех, и в группе студенты у всех. И получается, что у нас условная вероятность встречи этого признака итого "есть мобильный телефон" во всех группах, в группе студентки и студенты, одинаковая и равна 100%. И по всей выборке тоже 100%. То есть это отношение у нас получается равно единице, логарифм равен нулю получается. Ну что это значит? Это означает, что мы получаем нулевое количество информации о том, студент это или студентка, узнав, что у него есть мобильный телефон. Вот смысл этого всего. То есть это вот мера количества информации, называется мера Харкевича. Сам Александр Харкевич называл её прагматическая мера целесообразности информации. Такое название, но оно было хорошим в то время, когда он её предложил в шестьдесят седьмом году, по-моему. Вот. А сейчас у нас проведён уже за это время очень детальный анализ понятия данные, информация и знания. И можно сказать так, что обоснованно причём, что Харкевич предложил, по сути дела, не меру количества данных или меру количества информации, а меру количества знания. Но это я сейчас чуть попозже обосную вам. Харкевич – это известный советский учёный, академик Академии наук СССР, основатель Института проблем информации, который сейчас носит его имя, который внёс большой вклад в разработку... в разработку...

(Шум микрофона) Ребят, у вас там микрофоны включены, и в общем, мешает. Денис, выключи микрофон.

Вот. Вопрос возникает такой: а зачем брать логарифм? Ну для того, чтобы если в признаке нет информации о принадлежности объекта с этим признаком к классу, то получался ноль. Для этого берётся логарифм. Ну это одно из объяснений. А есть и другие способы тоже так сделать, чтобы получился ноль? Есть. Как вот я вам сейчас рассказывал, что можно взять отношение этих вот величин условного, безусловного, ну вероятности, можно говорить, и просто отнять единицу. Если они одинаковые, то получится ноль. Ну то есть этот коэффициент, он называется return on investment, investment. Значит, это возврат инвестиций, коэффициент возврата инвестиций. Если вверху у нас возвращённые инвестиции, внизу вложенные в объект инвестирования, то у нас получается так, что если мы получаем в результате от этого объекта инвестирования возврат больше, чем мы туда вложили, то тогда это рентабельно. Вот. Ну этот коэффициент отражает уровень рентабельности инвестиций, целесообразности.

Почему каждая формула два раза встречается? По той причине, что при расчётах вот этих, здесь используется n житое, которое представляет собой сумму по колонке, по житой колонке, соответствующей классу. Вот. И она может быть либо суммарным числом признаков, либо суммарным числом объектов обучающей выборки по этому, принадлежащих к этому классу. Это мера Хи-квадрат. Ну не сама мера, а один элемент этой меры. Вот. Это разница между фактической частотой наблюдения итого признака у объектов житого класса и теоретической частотой наблюдения этого признака у объектов, этого итого признака у объектов житого класса, где n - это число наблюдений всех признаков по всей выборке, n житое - это сумма по колонке, n итая - сумма по строке. Это замечательная мера, ребята, Хи-квадрат, замечательная мера причинно-следственных связей. Про Рои я вам рассказал. Есть ещё одна известная мне мера взаимосвязи, причинно-следственной взаимосвязи между наличием этого признака и принадлежностью объекта к житому классу – это просто разность условных и безусловных вероятностей, скажем так, относительных частот.

Теперь посмотрим, какая взаимосвязь между этими мерами, ребята. Взаимосвязь такая интересная. Смотрите, если мы возьмём этот логарифм и развернём его, то логарифм отношения, он равен разности логарифмов числителя и знаменателя. Получается логарифм P и житое минус логарифм P итое. Что нам это напоминает? Нам это напоминает сравнение этих относительных частот путём вычитания. То есть мы сравнили путём вычитания, а здесь мы сравнили путём деления. Они связаны между собой. Видно, да? Здесь тоже взяли путём отношения, вычисления отношения, но взяли минус единицу для того, чтобы нормировать к нулю, если зависимости нет.

А теперь давайте посмотрим вот на Хи-квадрат. И посмотрим на количество информации. Обратите внимание, в этом выражении, где через частоты написано количество информации Харкевича, здесь у нас в этих, в этом выражении под логарифмом есть такой элемент: n и житое умножить на n житое и n разделить на это произведение. Да? А если мы возьмём обратную величину, то что получится? n итая, n житая разделить на n. Видите, вот она. А здесь вверху n житая. Вот. То есть что получается? Что если мы вычисляем Хи-квадрат, то мы сравниваем частоты путём вычитания фактическую и теоретическую. А если мы возьмём логарифм, то мы их сравниваем путём деления. Мы мы частоту n и житое разделили на теоретическую частоту фактически. Видите, то есть умножили на n, разделили на n итая на n житая. То есть мы разделили вот на эту теоретическую частоту. Вам видно вот эту n итая умножить на n житая разделить на n - это теоретическая частота наблюдения итого признака у объектов житого класса. Вы меня понимаете, ребята? Улавливаете? То есть мы видим, что если мы возьмём разность этих величин, получается Хи-квадрат, а если возьмём отношение, логарифм отношения, то получается количество информации. То есть я хочу вам сказать, вы должны это понять, что количество информации по Харкевичу – это другой способ сравнения фактической и теоретической частот, чем Хи-квадрат. Если Хи-квадрат – это сравнение путём вычитания, то количество информации Харкевича – это сравнение путём деления. И всё, понимаете? Вот такая взаимосвязь между ними.

Надо где-то здесь это написать, наверное. (Пишет на экране) Сравнение частот вычитанием - Хи-квадрат. Сравнение частот делением - количество информации Харкевича.

Вот, это я напишу, потому что это существенный момент для будущих поколений. И вам тоже это объяснил.

**III. Система Эйдос**

**A. Обзор и особенности**

Таким образом, мы рассчитываем эти элементы и и житое. И рассчитываем все эти элементы и рассчитываем матрицу информативности. А в общем случае я называю эту матрицу системно-когнитивной модели. В системе Эйдос рассчитывается семь таких системно-когнитивных моделей. Они называются, имеют название: Инф1, Инф2, Инф3. А вот эти модели относительных частот и абсолютных называются ПРЦ1, ПРЦ2. От слова "проценты". Матрица абсолютных частот моделью АПС. Вот все эти модели рассчитываются в системе Эйдос.

То есть мы здесь видим, ребята, прямо количество информации. Вот слушайте дальше, уже здесь мы можем сказать. Это-то и есть количество информации. Здесь вот оно содержится это количество информации в матрицах относительных частот и абсолютных. Вот, она там содержится, но в явном виде её там нет, но она уже как бы просвечивает. То есть она там появляется путём сравнения вот этих условных, безусловных процентных распределений, которые мы производим в уме с помощью естественного интеллекта, так сказать. Она там как бы уже, ну, видна невооружённым глазом эта информация. А здесь она прямо вот непосредственно в виде представлена, вот прямо числами. Прямо будет здесь написано количество информации в битах, которое мы получаем, узнав, что у итого объекта житый, то есть и у некоторого объекта есть итый признак. Это количество информации о том, что, которое содержится в этом признаке о том, о факте наблюдения этого признака, о том, что этот объект относится к житому классу. То есть И и житое – это количество информации, которое мы получаем о принадлежности объекта с сытым признаком к житому классу.

И у нас так есть по этому признаку о всех классах. Значит, если мы получим, э-э, для по этому примеру посчитаем в уме студента или студентку, у нас здесь будет большое положительное количество информации в признаке "длинные волосы" о том, что это студентка. Что это студентка. Почему? Потому что это, э-э, если мы посмотрим на формулу, то это P и житое равно 99% всех студенток имеют длинные волосы, а в среднем 50% студентов имеют. То есть это логарифм двух. То есть мы получаем один бит информации. А для студентов, для студентов, ребят, мы получаем большую отрицательную величину. Вот здесь получается логарифм от величины, которая намного меньше единицы. Ну, скажем, единица, разделённая на 50, это одна сотая там, да? То есть получается логарифм одной сотой – это большая отрицательная величина. Что это, как это интерпретируется? Если мы знаем, что есть длинные волосы у того, кто там за дверью, то мы получаем из этого, зная этот признак, мы получаем большое количество информации о том, что это студентка, и большое количество информации о том, что это не студент. Вот это "не" – это минус.

Вот. Если будет признак итый – это наличие мобильного телефона, то здесь везде нули будут. С точностью расчётов до последнего знака, там может быть там погрешность в последнем знаке, потому что логарифм берётся. Вот. А если мы возьмём признак "брюки", например, "брюки". Значит, этот признак он у всех студентов наблюдается и почти у всех студенток. Получается, что у студентов он встречается чуть-чуть чаще, чем в среднем. Ну, допустим, 1,1. А у студенток он наблюдается 0,99 в среднем, а в среднем 0,98. Поняли, да? Вот. Нет, у студенток 0,98, а в среднем 0,99. И вот если мы разделим 0,99, то есть единицу разделим на 0,99, получится величина чуть-чуть больше единицы. А если мы разделим 0,98 на 0,99, то получим величину чуть-чуть меньше единицы. Получается, что в этом признаке "наличие брюк" у информации очень мало о том, что это студент. Потому что незначительным является отличие вероятности встречи этого признака в этой группе от среднего по всей выборке. И очень мало информации о том, что это не студентка. То есть здесь будет маленькая величина отрицательная, близкая к нулю, а здесь близкая к нулю положительная.

Ну теперь я очень важное понятие хочу вам объяснить. Два даже понятия. Это степень характерности признака для объектов класса. И второе – это ценность признака для решения задач идентификации, прогнозирования и других задач, которые можно решать с помощью таких моделей. Степень характерности признака. Если мы возьмём некий житый класс, ну, допустим, класс "студентки", и рассортируем все вот эти признаки, как в Экселе, целиком строку рассортируем, в порядке убывания количества информации в признаке о том, что объект принадлежит к этому классу "студентки". То у нас на первой позиции получится "длинные волосы" признак. Первая строчка будет в списке в этом ранжированном. Второй признак будет там "наличие телефона", и третий признак будет "наличие брюк". Причём наличие брюк будет с минусом, телефона с нулём, а длинные волосы с большим плюсом. Вот. То есть мы получили характеристику этого класса: какие признаки для него характерны, какие менее характерны, а какие вообще не характерны, а характерны для других классов.

А теперь ценность признака для решения задачи, допустим, распознавания, идентификации. Представьте себе, вот это признак у нас "длинные волосы". Тогда много информации о том, что это студентки, и много информации о том, что это не студент. Если мы посмотрим на вариабельность, то есть построим график по этой строчке количества информации, то этот график у нас будет, что характерно, что для него будет? Что у него будет большая амплитуда положительная и большая отрицательная. То есть он будет сильно отклоняться от нуля или от среднего. А если мы возьмём признак "брюки", вот, он будет мало отклоняться от среднего. У него будет мало информации о том, что это студент, и мало информации о том, что это не студентка. Он будет мало отклоняться от нуля. А если возьмём признак "наличие телефона", то он вообще просто не будет отклоняться от нуля. Можно ли оценить, как можно интерпретировать эту вот вариабельность количества информации в строке? Я её интерпретирую как ценность признака для решения задач идентификации, прогнозирования, принятия решений и других задач. То есть можно сказать так, что если мы все признаки, для всех признаков посчитаем вот эту величину... Да, как можно определить количественно эту вариабельность? Можно разными способами. Ну, например, можно взять, вычислить среднее по строке, а потом взять и модули, там же плюсы и минусы есть, положительные числа и отрицательные. Поэтому нужно брать модули. Модуль отклонения от средних, от среднего. Сумма модулей отклонения от среднего. А можно взять среднее квадратичное отклонение посчитать вот таким способом. В общем, есть различные меры вариабельности. Не брать квадратный корень, будет дисперсия. Вот. Ну, в общем, понятно, что есть много разных мер вариабельности. Наиболее такой привычной является стандартное отклонение, среднеквадратичное, лучше называть, чтобы не связываться с распределением Гаусса. И получается что? Что мы можем для каждого признака посчитать эту вариабельность, а потом их ранжировать в порядке убывания вариабельности. Тогда у нас получится сначала самые ценные для того, чтобы отличить один объект от другого признаки, потом менее ценные, а в самом конце будут вообще бесполезные признаки.

Вот то, что я вам хотел рассказать про модели. Вот. Теперь, следующий вопрос у нас... Вы экран видите, да? Следующий вопрос – это интегральные критерии. Значит, у нас есть задача определения степени сходства, задача идентификации, определение степени сходства некоторого объекта или явления, или состояния объекта, типа объекта, или явления, процесса, с определёнными обобщёнными образами, которые у нас сформированы в модели этих объектов, явлений и процессов. Вот, допустим, у нас есть некий объект, у которого есть определённый набор признаков. Некоторые признаки у него наблюдаются, а другие не наблюдаются. Что мы можем сказать об этом объекте, насколько он сходен с тем или иным классом? Как определить степень сходства объекта с обобщённым, конкретного объекта с обобщённым образом класса? Используя математику, то есть математические вот эти модели, которые сейчас я вам рассказываю. Для этого в системе Эйдос, вообще применяется такая технология. Значит, представьте себе, значит, я просто сейчас скажу словами. Мы будем считать, что объект принадлежит к различным классам в различной степени. Ну то есть он на один класс больше похож, на другой меньше, а на третий вообще не похож. Но если мы не будем рассматривать целый список классов, а у нас стоит задача определить, к какому классу он относится, а к одному, да? То тогда можно сказать так: мы можем предположить... Сейчас слушайте внимательно. Мы можем предположить, что объект относится к тому классу, к той обобщающей, обобщённой группе объектов, о принадлежности к которой в его признаках наибольшее количество информации. То есть мы знаем, какое количество информации содержится в каждом признаке о принадлежности или непринадлежности объекта с этим признаком к каждому из классов. О каждом признаке мы это знаем. А если признаков много у объекта? Ну тогда напрашивается мысль просуммировать это количество информации. Именно просуммировать. Раз просуммировать, это называется аддитивный интегральный критерий, в отличие от мультипликативного, где перемножаются частные критерии. Но если мы рассматриваем частные критерии как количество информации, а у нас именно так они и рассматриваются, вот это количество информации по Харкевичу. Ну а другие – это не количество информации, но связанные с ним, такие как Хи-квадрат, Рои. Вот. То напрашивается такая мысль: описать каким-то образом объект сам идентифицируемый, описать, какие у него есть признаки, какие есть, какие нет. И просто сравнивать, рассчитывать, рассчитывать суммарное количество информации в признаках объекта о принадлежности к каждому из классов по очереди. А потом их просто проранжировать в порядке убывания этого количества информации. Вот такая вот идея в основе лежит этого подхода.

Можно суммировать как? Организовать цикл по признакам. Если признак есть, суммировать количество информации о принадлежности к житому классу. Если нет признака, то не суммировать. А до этого организовать цикл по классам. То есть мы для каждого класса рассчитываем суммарное количество информации в признаках объекта о принадлежности к этому классу. Но для того, чтобы избавиться от сравнения "если есть, то суммировать, если нет, то не суммировать", как можно сделать, чтобы всегда суммировать, ребята? Ну я подумал, что лучше от логики освободиться, сделать что-то такое, что можно формулами записать. Ну я думаю, что это сделать очень просто. Для этого нужно просто, когда признак есть, то умножать на количество информации единичку, а когда признака нет, то нолик умножать на количество информации. То есть речь идёт о чём? О том, что вот это И и житое, ребят, это суммарное количество информации в признаках итого, то есть в признаках объекта, признаках некоторого объекта, о том, что он принадлежит к житому классу. Значит, для этого нужно описать объект массивом или вектором таким, у которого такие координаты: если признак есть, тогда единичка, если признака нету, тогда нолик элемент массива этого. Вот. И таким образом при суммировании вот так получится, что мы умножаем на количество информации только те, в том случае, если признак есть, и суммируем. Ну то есть, по сути дела, эта формула – это формула суммы количества информации в признаках объекта, которые у него есть.

Значит, здесь вот ещё есть такая средняя строчка, что может быть не единичка, а n, если признак наблюдается n раз. Что это значит? Ну, например, если мы берём слово "молоко" и идентифицируем его, в качестве признаков рассматриваем буквы, то буква О три раза встречается. То есть мы здесь тогда используем не первый вариант, что этот признак есть, а он есть, но он есть три раза, понимаете? Тогда мы вместо единички суммируем три, умножаем на три. Ну вот, собственно, так. Теперь, если мы на эту формулу посмотрим и вспомним немножко линейную алгебру, аналитическую геометрию, я не знаю, у вас была или нет аналитическая геометрия, ребят, скажите, была у вас или нет? Ну, векторная... Короче, вы знаете, что такое произведение векторов? Скалярное, векторное произведение, знаете, нет? Короче говоря, вот эта вот формула, она представляет собой формулу скалярного произведения в координатной форме. Это вот векторная форма скалярного произведения из линейной алгебры, а это вот в координатной форме. Произведение этих двух векторов. Один вектор объекта, а другой вектор класса житого. Вот. Объект описан вектором, где у него нули, единички координаты. Признак есть – единичка, нету – нолик. А класс описан количеством информации в этом признаке о принадлежности к житому классу. Вот вектор класса. И эти два вектора перемножаются скалярно и получается у нас интегральный критерий очень интересный. У него, я считаю, есть две очень важных особенности. Первая важная особенность: он применим независимо от того, является ли вот это пространство, определённое вот этой матрицей, ортонормированным. Дело в том, что вот, скажем, мера Эвклида, она корректна только для ортонормированных пространств. А вот эта мера, векторное расстояние, можно его назвать, или информационное расстояние, оно, ей совершенно всё равно, какое там пространство. Что значит ортонормированное или неортонормированное? Это значит, классы вот эти вот, вектора классов, вот эти, вектора классов, они коррелируют друг с другом или нет? Если пространство ортонормированное, ребят, то корреляция вот этих вот векторов классов равна нулю. Они взаимно перпендикулярны все, понимаете? Вот. То есть каждый из них – это ось в некотором пространстве. И они, эти оси взаимно перпендикулярны. Значит, ну я могу вам сказать, что даже и так интуитивно понятно, что вообще говоря, то есть в общем случае, они не взаимно перпендикулярны. То есть могут быть классы, которые похожи друг на друга. Вот. То есть вообще говоря, почему бы и нет? Могут быть такие данные, что они будут похожи друг на друга. Тогда пространство будет неортонормированное. И тогда мера расстояния Эвклида, теорема Пифагора, будет неприменима. А вот это вот векторное расстояние, информационное, оно, ей совершенно всё равно, какое там пространство. Это первое. Второе. Если мы посмотрим на эту формулу, то по сути это определение шума. Если эта сумма равна нулю, то это, если мы возьмём один элемент сигнала, какое-то множество сигнала и другое подмножество сигнала, L - это вектор L - это одно подмножество, а И и житое - другое подмножество. То есть мы берём, рассматриваем эти вектора, вектор класса и вектор объекта, который вот здесь нет, не приведено, но аналогичный вектор, только там нули единички. Мы их рассматриваем как различные элементы сигнала, идущие один за одним. И нас интересует, связаны ли будущие элементы сигнала с прошлыми элементами сигнала, значениями. Если они не связаны, тогда этот сигнал называется шум. То есть это определение шума математическое, белого шума. Из этого вытекает очень интересный вывод. Если у нас будут случайным образом связаны признаки с объектами, то есть эти вот признаки, итые признаки будут случайным образом связаны с классами, ну то есть объекты, которые предъявлены в результате наблюдений, обучающая выборка будет содержать случайные там числа просто-напросто вот. То тогда вот эта величина, она будет, как правило, равна близка к нулю. То есть мы увидим, что это шум. А реальный сигнал, а дальше слушайте внимательно, реальный сигнал, он всегда является суммой истинного сигнала, который мы хотим исследовать, и шума какого-то, который вкрался в этот истинный сигнал при наблюдениях, измерениях там и тому подобное, каких-то операциях с ним. Поэтому такой интегральный критерий, он, ну, можно так сказать, он этот шум игнорирует, который есть в исходных данных. То есть эта модель устойчива к шуму.

**B. Сравнение с другими системами**

Теперь, если мы возьмём и заменим каждый из векторов стандартизированными значениями координат, то есть его координаты заменим стандартизированными значениями. У нас тогда получится... Вот что-то я их не нарисовал эти стандартизированные значения. Вот. Ну, стандартизированное значение как вычисляется? Берётся значение элемента вектора, вычитается среднее из него, вот, и делится на стандартное отклонение или среднеквадратичное отклонение значений этой координаты этого вектора. То есть значение заменяется его стандартизированным значением. Значение координат вектора заменяется стандартизированным значением координат. Есть и другие способы нормировки. Например, можно значение, из значения вычесть минимальное значение из текущего значения и разделить на разницу между максимальным и минимальными значениями. Тоже будет нормировка, похожая вот на стандартизацию. Мы получаем вот такой интегральный критерий. Он похож очень на вот этот, который был до этого, скалярное произведение. Разница в чём? Разница в том, что здесь не нормированные вектора, как они были, так они здесь и используются. А здесь нормированные вектора, стандартизированные.

**C. Ключевые этапы анализа в Эйдос**

Ну и всё, больше разницы никакой нет. Это я рассказал вам про интегральные критерии. Вот. Теперь, соответственно, мы можем решать задачи, используя эти модели и рассказанные, описанные мной интегральные критерии. Мы можем решать задачи идентификации, принятия решений и исследования моделируемой предметной области путём исследования модели этой области.

На этом мы закончили тему, которую мы начали на прошлом, на прошлой лекции. Тема первая глава, то есть в пособии: данные, информация, знания, определение содержания понятий. И в более подробном варианте я дал не только определение этих понятий и их соотношение такое на уровне логики, но и описал математическую модель преобразования данных в информацию, а её в знание, конкретную математическую модель. Что об этой модели можно интересного сказать, ребята? Что эта модель является универсальной, не зависящей от предметной области, и может применяться в самых различных предметных областях успешно. И чем я и занимался. То есть я и применял её в самых различных предметных областях. Вот. Сейчас я вам покажу. Это мой сайт. Вспоминаете, нет? Кто вспоминает?

И мы смотрим второй пункт. Какая тишина, как будто никто не вспоминает, что это мой сайт. Второй пункт. Там выбираем опцию "скачать и запустить систему Эйдос" и листаем вот так. Тут много всего. Структура есть, как её скачать, установить, структура системы, монография, свидетельство Роспатента. А потом тематические подборки публикаций, смотрите: по информационным мерам сложности, анализу изображений, анализу текста, когнитивным функциям, логике, методологии, методологии познания, изучению влияния космической среды на различные процессы на Земле, современным интеллектуальным коммуникационным технологиям, виртуальной реальности, когнитивной ветеринарии, тематика связанная с АПК. То есть это подборки прямо публикаций, которые вы можете посмотреть любую из этих публикаций. Вот. То есть эти модели широко применялись в самых разных предметных областях и очень успешно. Получались хорошие результаты. Это просто статьи в открытом доступе. Здесь вот 244 статьи на данный момент. Это статьи в журнале в электронном.

Вот. Теперь у нас следующий вопрос, ребята. Значит, я назову дисциплину: Интеллектуальные информационные системы и технология. 11 сентября. Рассматриваем вопрос, который тема во второй главе: выявление, представление и использование знаний для решения задач идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области путём исследования модели этой предметной области. И эта глава, она посвящена рассмотрению конкретного примера. Я предлагаю рассмотреть решение этих задач на конкретном примере. Прямо взять конкретный пример и посмотреть. Для в качестве такого примера взята статья моя, которую, ссылочку на неё здесь почему-то я не привёл. Посмотрим попозже. Там должна быть ссылка. Ага, вот. Может, это она и есть эта ссылка?

Вот. Теоретические основы системно-когнитивного моделирования процессов и машин агроинженерных систем. То есть это когнитивная механизация, процессы и машины агроинженерных систем, механизация. И вот в этой статье подробно-подробно расписано: суть метода, а потом написано, видите, когнитивно-целевая структуризация, формализация, синтез, верификация, повышение качества, выбор наиболее достоверной модели, решение задач идентификации, классификации, принятия решений, исследования моделируемой предметной области.

**IV. Заключение и следующие шаги**

Значит, э-э, я считаю, украшают эту систему, скажем так, являются её достоинствами. Она разработана в универсальной постановке, не зависящей от предметной области, и может быть применена во многих предметных областях. Вот, то есть в экономике, медицине, технических науках, там, ветеринарии, где, психологии, медицине, в общем, во многом областях, многих областях. Значит, очень интересным моментом является то, что система находится в полном открытом бесплатном доступе. Причём с актуальными исходными текстами. Значит, вот исходный текст системы прямо в интернете, видите, он находится. Это для англоязычных написано то же самое, что я сейчас вам рассказывал, что она делает. Вот. Сейчас текущая версия является версия конца июля. Потом я там писал пособие. Сейчас написал летом его в августе, когда отпуск был. Вот. Ну, значит, я возьму немножко так для прикола полистаю. Полистаю этот текст. Вот. Значит, ну, значит, давайте полистаю, он загружается. Вот. Ну, в общем, этот текст очень большой, ребят. То есть это так вот, я держу на Page Down руку, даже не на стрелочке вниз, а Page Down. А справа там бежит бегуночек, и вы, наверное, видите, как он бежит. Ну, в общем, у системы довольно большой исходный текст. Около, больше 3.000 страниц. Нет, извините, больше, да, э-э, э-э, больше 3.000 листов. Примерно 127, там, 128.000 строк. То есть три, то есть больше 3.000 листов. То есть он очень большой. 100% этого исходного кода, который здесь вот я показал вам, написано мною. То есть никто больше исходный код на языке XBase++, на котором написана система Эйдос, не писал. То есть только я его писал. Это полностью написано мной. Ну, конечно, я использовал технические решения, которые есть в литературе. Вот, и на сайтах разработчиков. Значит, вот сайт разработчиков. Здесь довольно часто я там что-то спрашивал, если было мне непонятно. Вот, скажем, Евгений Луценко, видите, он есть. А это я пояснял наоборот, не только спрашивал, но и помогал, подсказывал. Вот.

Систем существует много, вот, но я могу сказать, что они достаточно сложные в использовании обычно. И поэтому я поставил себе цель разработать простую систему, такую, которая была бы проста для пользователей. И могу вам сказать, что эту цель сейчас и фирма Microsoft поставила перед собой и разработала два варианта интеллектуальных систем в открытом доступе с нулевым порогом входа практически. То есть простых в освоении. Но дело в том, что это сейчас произошло, а системе Эйдос уже больше 30 лет. Потому что, ну, даже таких там тридцати. Если первая её версия была в восемьдесят первом году уже запущена в эксплуатацию, то есть это там 20 лет и здесь 19. Ну вот почти 40 лет уже это развивается. То есть я это постоянно этот проект развиваю, совершенствую. Огромное количество расчётов проведено, описано в разных статьях, книгах, в большом количестве. И защищено довольно много диссертаций. Я вам, по-моему, даже и говорил. Пять экономических докторских диссертаций, два, две по техническим наукам докторских, одна по биологии, четыре по психологии, одна по техническим наукам кандидатская, одна по экономике, одна по медицине. Ну то есть много защищено диссертаций с использованием этой технологии в исследованиях.

Так вот, ребята, значит, какие положительные особенности есть у системы Эйдос? Чем она, так сказать, интересна? Тем, что она разработана в универсальной постановке, не зависящей от предметной области, и может быть применена во многих предметных областях. Вот. То есть в экономике, медицине, технических науках, там, ветеринарии, где, психологии, медицине, в общем, во многом областях, многих областях. Значит, очень интересным моментом является то, что система находится в полном открытом бесплатном доступе. Причём с актуальными исходными текстами. Значит, вот исходный текст системы прямо в интернете, видите, он находится. Это для англоязычных написано то же самое, что я сейчас вам рассказывал, что она делает. Вот. Сейчас текущая версия является версия конца июля. Потом я там писал пособие. Сейчас написал летом его в августе, когда отпуск был. Вот. Ну, значит, я возьму немножко так для прикола полистаю. Полистаю этот текст. Вот. Значит, ну, значит, давайте полистаю, он загружается. Вот. Ну, в общем, этот текст очень большой, ребят. То есть это так вот, я держу на Page Down руку, даже не на стрелочке вниз, а Page Down. А справа там бежит бегуночек, и вы, наверное, видите, как он бежит. Ну, в общем, у системы довольно большой исходный текст. Около, больше 3.000 страниц. Нет, извините, больше, да, э-э, э-э, больше 3.000 листов. Примерно 127, там, 128.000 строк. То есть три, то есть больше 3.000 листов. То есть он очень большой. 100% этого исходного кода, который здесь вот я показал вам, написано мною. То есть никто больше исходный код на языке XBase++, на котором написана система Эйдос, не писал. То есть только я его писал. Это полностью написано мной. Ну, конечно, я использовал технические решения, которые есть в литературе. Вот, и на сайтах разработчиков. Значит, вот сайт разработчиков. Здесь довольно часто я там что-то спрашивал, если было мне непонятно. Вот, скажем, Евгений Луценко, видите, он есть. А это я пояснял наоборот, не только спрашивал, но и помогал, подсказывал. Вот.

Она является одной из первых систем отечественных искусственного интеллекта. Ну я могу вам показать, что вот это внедрение восемьдесят седьмого года. Вот смотрите, центр аэрокосмический Кубанский. Значит, была разработана система, было процентное распределение, а также распределение информативности признаков. То есть уже эта модель использовалась. И вот, видите, здесь внизу моя подпись: главный конструктор проекта Луценко, восемьдесят седьмой год. То есть уже тогда эта система позволяла рассчитывать позитивные и негативные информационные портреты. Видите? То есть она решала задачу создания моделей вот этих, которые я вам показывал. А информационный портрет – это просто выборка признаков определённого класса и сортировка в порядке убывания. Какие признаки наиболее характерны, какие наиболее не характерны для того или иного класса. Так это было в восемьдесят седьмом году, ребята. Понимаете? Я в этом центре работал сначала начальником отдела, а потом главным конструктором стал.

И это всё было реализовано в среде системы Вега М, моей разработки. Работа была выполнена с Академией наук по заказу Академии наук. То есть эта система, она тогда уже работала на компьютерах ВАНГ советского производства. Компьютеры ВАНГ советского производства назывались Искра 226.

Она обеспечивает устойчивое выявление в сопоставимой форме силы и направления причинно-следственных зависимостей, неполных, зашумлённых, взаимозависимых и линейных данных очень большой размерности, числовой и нечисловой природы, измеряемых в различных типах шкал: номинальных, порядковых и числовых, в различных единицах измерения. То есть не предъявляя жёстких требований к данным, которые невозможно выполнить, и обрабатывает те данные, которые есть. Это очень важное достоинство, уникальное для этой системы, которое не обеспечивается ни системой SPSS, ни системой Статистика, ни системой Дедактор.

Содержит большое количество локальных, поставляемых с инсталляцией, облачных Эйдос-приложений. В настоящее время локальных приложений 31, поставляемых в инсталляции, и 207, ребята, облачных Эйдос-приложений. Сейчас покажу вам их. Да, здесь ссылка неправильная. Исправлю.

Вот. Здесь описана среда Эйдос описана. Она имеет мультиязычную поддержку интерфейса. В настоящее время она поддерживает интерфейс на пятидесяти одном языке. То есть на русском языке и ещё на пятидесяти языках. Языковые базы могут генерироваться пользователями, а поставляются и в инсталляции, в полной инсталляции. Поддерживает онлайн-среду накопления знания и широко используется во всём мире. Вот это вот интересное. Ну я так думаю, благодаря тому, что она бесплатная, это первое. И второе, благодаря тому, что она работает. То есть она реально решает задачи. Вот. Представляет собой интерес для людей, которые не хотят особо там ломать голову над тем, как это всё делать. Потому что система имеет нулевой порог входа. То есть там все хелпы контекстно-зависимые, масса примеров решения задач учебных, встроенных и внешних. Ну это вот карта кластеров её применения. Мест, где она запускается. Ну видно, что её запускают и в России, и в Европе, и в Китае, и в США. Вот, например, там довольно много посещений, поэтому она дурит, лагает, наверное, из-за него. Вот, США и Канада. Видите? Особенно вот в США много запусков в Калифорнии, на западе США её запускают очень интенсивно. Вот. Здесь, если мы посмотрим, ещё этот Webex сейчас перетормаживает, наверное, из-за него. Хорошо так. Ну, в общем, в Калифорнии, на западе США её запускают очень интенсивно.

Наиболее трудоёмкие в вычислительном отношении операции синтеза моделей и распознавания реализованы с помощью графического процессора ГПУ, что на некоторых задачах обеспечивает ускорение решения этих задач в 1000 раз, в 1000 раз. У меня была задача, ну на разных задачах по-разному ускорение, там в зависимости от того, какие операции там наиболее часто встречаются при решении, при решении этих задач. Но я как бы зафиксировал, когда в 4.000 раз быстрее она сработала, чем на центральном процессоре. Что обеспечивает реальную, ребята, реальную обработку больших данных Big Data, а также большой информации и больших знаний. Ну вот большая информация, большие знания – это термины ввёл я, их нету, не использовались они. Вот. То есть что такое 4.000 раз, ребят? Ну я так для примера вам скажу, в году сколько дней? 300 там 65, да, будем считать. В десяти годах сколько дней? 3.600 там 50 дней, да? То есть меньше 4.000 дней. Так вот представьте себе, что мы считаем день на графическом процессоре задачу, даже меньше, там 20 часов там или, скажем, 18 часов. Или считаем 10 лет на центральном процессоре, на компьютере i7, кстати, с 16 ГБ оперативки. Можете себе представить, что это такое? Это принципиально. Одно дело около суток, другое дело 10 лет, правильно? Вот.

Обеспечивает преобразование исходных эмпирических данных в информацию, а её в знание и решение с использованием этих знаний задач классификации, поддержки принятия решений, исследования предметной области путём исследования её системно-когнитивной модели, генерирует при этом очень большое количество табличных и графических выходных форм. То есть у неё развита когнитивная графика в этой системе. Подавляющее количество этих выходных форм графических не имеет никаких аналогов в других системах, таких как Excel там или StatGraph. Вот. И примеры таких форм можно посмотреть вот в этой работе, на которую ссылка даётся. Хорошо имитирует человеческий стиль мышления. То есть эксперты, рассматривая результаты решения в системе, как естественные, очевидные. Они говорят: "Да, мы это знаем, это нам известно". Я говорю: "Вам-то известно, а нам-то откуда известно?" Мы взяли ваши журналы агрономов обработали и выявили законы севооборота, понимаете, так, которые вообще-то, я не знаю, там за тысячелетний опыт обобщены в этих законах. Причём выявили их для нашей конкретной местности, конкретного времени. И вот здесь вот я ещё раз почётче сформулировал, что вместо того, чтобы предъявлять к исходным данным практически неосуществимые требования, вроде нормальности распределения, абсолютной точности, полноты и повторности всех сочетаний значений факторов и их полной независимости, аддитивности, автоматизированный системно-когнитивный анализ предлагает без каких-либо предварительной обработки осмыслить эти данные, тем самым преобразовать их в информацию, а затем преобразовать эту информацию в знание путём её применения для достижения целей, то есть для управления и решения задач классификации, поддержки принятия решений и содержательного эмпирического исследования моделируемой предметной области. Эффективность подхода, реализованного в системе, не зависит от того, что мы думаем о предметной области, и думаем ли вообще, знаем ли мы её. Поэтому именно эти модели эффективны даже тогда, когда наши представления о предметной области ошибочны или отсутствуют. Вот. Модели, которые предлагает система – это феноменологические модели. То есть они отражают только сам факт детерминации, силы и направления причинно-следственных связей, но не отражают механизмов этой детерминации, каким образом получается, что именно вот такие причинно-следственные связи.

Ребят, когда будет приближаться конец занятия, вы мне напомните. И, значит, э-э, про систему я вам коротко рассказал. Теперь, э-э, давайте, э-э, я сформулирую, э-э, как, э-э, основные этапы технологические постановки и решения задач в этой системе. Прямо перечислю их. Первый этап называется когнитивная структуризация предметной области. Иногда ему называют когнитивно-целевая структуризация предметной области. Это единственный неавтоматизированный этап системно-когнитивного анализа. Все остальные этапы автоматизированы. Второй этап... Значит, что мы на этом этапе делаем на этапе когнитивной структуризации предметной области? Мы определяем, э-э, я бы сказал так, мы формулируем идею приложения или идею решаемой задачи. Мы, эта идея формулируется в очень обобщённой форме, на таком уровне примерно: что мы хотим рассматривать как причины, как факторы, влияющие на объект моделирования, и что мы хотим или будем рассматривать в качестве результатов влияния этих факторов? Ну, обычно это какие-то состояния объекта моделирования, текущие и будущие. То есть что мы рассматриваем в качестве причин, влияющих на объект моделирования, и что мы рассматриваем в результате действия этих причин? Вот это определяется на этом этапе, на интуитивном уровне. Человек понимает, что вот его интересует, как влияет вот это на вот это. Вот система Эйдос, она позволяет выявлять, что, как влияет и на что. Что и на что, и как. Положительно, отрицательно, с какой силой, способствует или препятствует, и с какой силой способствует или препятствует.

Значит, ну на таком уровне, когда мы задачу формулируем, то тогда ещё не удаётся как-то её решить в системе программной на компьютере. Для этого необходимо повысить степень формализации описания задачи. Этот этап называется формализация предметной области. Когда я вам рассказывал про преобразование данных в информацию, а её в знание, то, по сути дела, я вам рассказывал о том, что делается на этом этапе. То есть исходные данные мы должны преобразовать в информацию. Для этого мы что делаем? Разрабатываем справочники классификационных и описательных шкал и градаций – это первое. Второе – мы эти справочники применяем для того, чтобы закодировать исходные данные и создать базу событий. Третье – мы эту базу событий анализируем с помощью различных мер причинно-следственных взаимосвязей, которые я вам показывал в табличке вот этой, таблица три, различных мер причинно-следственных связей. И потом берём и рассчитываем... Да, и формируем, значит, обучающую выборку, то есть эти базу событий, а потом используем её для того, чтобы рассчитать причинно-следственные связи между событиями. На этом у нас этот этап заканчивается. Вот. Даже вот у нас ещё синтез и верификация модели. То есть у нас этап формализация предполагает разработку шкал и обучающей выборки, и всё. То есть мы ещё в информацию не преобразовали исходные данные. Завершается преобразование в информацию, когда мы осуществляем синтез и верификацию моделей. Ну, синтез – это расчёт, а верификация – это проверка их на достоверность, оценка их достоверности. Это тоже важный очень и интересный вопрос, как оценивать достоверность моделей. Мы его рассмотрим обязательно подробно. Потом у нас есть различные способы повышения качества модели, повышения достоверности. И мы эти способы можем применить, можем не применить – это наше дело. А я вам скажу так, обычно не требуется их применять. То есть модель имеет достаточно высокую достоверность и без этого. Но, в принципе, можно это заняться повышением достоверности моделей. Ну это отдельный этап работы. Потом мы из всех созданных моделей выбираем наиболее достоверную. Я могу вам сказать, ребят, чтобы вы уже это, наверное, поняли, что за один расчёт, когда мы осуществляем синтез и верификацию моделей, сразу создаётся 10 моделей в системе Эйдос. Три модели статистические – это модель корреляционная матрица, модели условных, безусловных процентных распределений, и семь моделей системно-когнитивных, связанных с информацией, количеством информации по Харкевичу, Хи-квадрат, коэффициент Рои и коэффициент взаимосвязи. А потом в этой наиболее достоверной модели, которую мы выбрали по критериям достоверности, мы решаем в ней задачи диагностики, классификации, распознавания, идентификации, задачи поддержки принятия решений и задачи исследования моделируемой предметной области путём исследования модели. Этих задач или подзадач, их тут довольно много. Здесь они не перечислены, здесь перечислено так укрупнённо всё это.

Теперь дальше предлагается что сделать? Значит, дальше у меня идёт здесь разделчик такой в пособии 2.1. Преобразование данных в информацию, нормализация баз исходных данных, формирование баз событий, выявление смысла в событиях, то есть выявление причинно-следственных связей. А потом в знание: классификация будущих состояний как целевых и нежелательных, оценка силы и направления влияния факторов на достижение целевых состояний.

У нас, похоже, дело идёт к тому, что сейчас конец занятия, да? Правильно?  
(Ответ: Да)  
Правильно. Значит, тогда попрошу старосту, Аню тоже, вот, сказать, что мы начали, только начали рассматривать раздел 2.1. То есть мы стали рассматривать конкретное применение этой технологии для решения конкретной задачи. Раздел 2.1 пособия. Ну и, в общем-то, почти что ничего не рассмотрели, только дошёл я до описания исходных данных. Вот это, пожалуйста, напомните мне на следующем занятии. Даже можно сказать страницу 36 страница, к примеру, вот. 36, 37 страница пособия.

Вот. Если у вас какие-то есть вопросы, то я думаю, что на перерыве не будем мы их задавать, а можете задать мне их либо в Мудле там в кабинете, либо, если вам дали пароль, а, Ань? Либо можете на следующем занятии с этого начать, что задать мне вопросы. Вот. Я какие у вас возникают вопросы? Потом вы можете спокойно задавать вопросы просто во время занятия. Мне даже это будет нравиться, потому что я вас всё время что-то спрашиваю, чтобы узнать, слушаете вы или не слушаете, там вы, не там. Так что можете спрашивать во время занятия. Всё, до свидания. Всего самого хорошего. Запись прекращается.

Так, если я останавливаю запись, то происходит формирование файла, да?