

## Тема 7. Методы кодирования социально-экономической информации. 2 часа

Для того чтобы классификационная схема стала средством для создания эффективной системы хранения и инструментом обработки информации, она должна быть дополнена системой условных обозначений, присваиваемых объектам. Такие обозначения в разных классификаторах называются индексами или кодами, а процесс их присвоения объектам классификации – индексированием или **кодированием**.

Знаки, составляющие индекс или код, называются **алфавитом**. Такой алфавит может включать буквы, цифры, знаки пунктуации в различных комбинациях. Код является идентификатором объекта классификации и его основное назначение состоит в однозначном обозначении объекта. Это своего рода формализованное имя объекта, которое должно обеспечивать возможность точного определения объекта классификации. Поэтому разработчики классификационных схем стремятся сделать индексы или коды мнемоничными, то есть такими, чтобы даже по внешнему виду, алфавиту кода пользователь мог определить объект и узнать возможно больше информации о характере объекта классификации, для обозначения которого использован этот код. Например, буквенный код России по Общероссийскому классификатору стран мира (ОКСМ) – RU, а код США – US. Соответственно код российского рубля по Общероссийскому классификатору валют (ОКВ) – RUR, а доллара США – USD.

Кодирование есть процесс преобразования одного алфавита сообщения в другой алфавит. Применительно к документам, кодирование можно рассматривать как процесс присвоения документу единственного обозначения – кода, которое отличает один документ от другого, т.е. идентифицирует документ. С другой стороны, кодирование документа можно рассматривать как процесс преобразования документа на естественном языке в язык кодов.

Основное назначение кодирования состоит в приспособлении информационного сообщения к каналу связи. Кодирование документов и документной

информации направлено на приспособление документа к возможностям его обработки с помощью средств вычислительной техники.

К методам кодирования ТЭСИ предъявляются определенные требования, соблюдение которых способствует повышению качества классификатора. Метод кодирования должен:

- предусматривать использование в качестве алфавита кода десятичных цифр и букв;
- обеспечивать по возможности минимальную длину кода и достаточный резерв незанятых позиций для кодирования новых объектов без нарушения структуры классификатора;
- быть максимально ориентированным на автоматизированную обработку информации.

Методы кодирования могут носить самостоятельный характер – регистрационные методы кодирования или быть основанными на предварительной классификации объектов – классификационные методы кодирования.

**Регистрационные** методы кодирования бывают двух видов: порядковый и серийно-порядковый. **Порядковый** метод кодирования – это такой метод, при котором кодами служат числа натурального ряда. В этом случае каждый из объектов классифицируемого множества кодируется путем присвоения ему текущего порядкового номера.

Преимущества:

1. Метод обеспечивает довольно большую долговечность классификатора при незначительной избыточности кода.
2. Простота метода. Используются наиболее короткие коды и лучше обеспечивается однозначность определения каждого объекта классификации.
3. Обеспечивает наиболее простое присвоение кодов новым объектам, появляющимся в процессе ведения классификатора.

Недостатки:

1. Отсутствие в коде какой-либо конкретной информации о свойствах объекта.

2. Нет возможности размещения вновь появившихся объектов классификации в необходимом месте классификатора, так как резервные коды располагаются в конце ряда.

По этим причинам порядковый метод кодирования отдельно очень редко применяется при создании классификаторов ТЭСИ. Чаше всего он применяется в сочетании с другими методами кодирования.

**Серийно-порядковый** метод кодирования – это такой метод, при котором кодами служат числа натурального ряда с закреплением отдельных серий этих чисел (интервалов натурального ряда) за объектами классификации с одинаковыми признаками. В каждой серии, кроме кодов имеющихся объектов классификации, предусматривается определенное количество кодов для резерва. Резерв кодов располагается в середине или в конце серии. Это является большим преимуществом данного метода по сравнению с порядковым методом кодирования. Серийно-порядковый метод кодирования целесообразно применять для объектов, имеющих два соподчиненных признака. Данный метод кодирования обладает всеми преимуществами и недостатками порядкового метода кодирования. Обычно этот метод используется для идентификации объектов в сочетании с классификационными методами кодирования.

**Классификационные** методы кодирования бывают двух видов: последовательный и параллельный. **Последовательный** метод кодирования – это такой метод, при котором код объекта классификации образуется с использованием кодов последовательно расположенных подчиненных группировок, полученных при иерархическом методе классификации. В этом случае код нижестоящей группировки образуется путем добавления соответствующего количества разрядов к коду вышестоящей группировки. Последовательный метод кодирования чаще всего используется при иерархическом методе классификации.

Преимущества последовательного метода кодирования являются логичность построения кода и большая емкость. Вместе с тем он обладает всеми недостатками, присущими иерархическому методу классификации, а также ограниченными возможностями идентификации объектов. Главный недостаток –

отсутствие гибкости. В результате зависимости значений последующих разрядов кода от предыдущих нельзя применять этот код по частям, невозможно вносить новые признаки и производить изменения в коде без коренной перестройки классификатора. Поэтому применять последовательный метод кодирования целесообразно в тех случаях, когда набор признаков классификации и их последовательность стабильны в течение длительного времени.

**Параллельный метод** кодирования – это метод, при котором код объекта классификации образуется с использованием кодов независимых группировок, полученных при фасетном методе классификации. При этом методе кодирования признаки объекта кодируются независимо друг от друга. Для параллельного метода кодирования возможны два варианта записи кодов объектов:

1. Каждый фасет и значение фасета имеют свои коды, которые включаются в состав кода объекта. Такой способ записи удобно применять тогда, когда объекты характеризуются неодинаковым набором признаков и различным их числом. При формировании кода какого-либо объекта берутся только необходимые признаки.

2. Для определенных групп объектов выделяется фиксированный набор признаков и устанавливается стабильный порядок их следования, то есть устанавливается фасетная формула. В этом случае не надо каждый раз указывать, значение какого признака приведено в определенных разрядах кода объекта.

Параллельный метод кодирования имеет ряд преимуществ.

- Гибкость структуры кода, обусловленная независимостью признаков.
- Метод позволяет использовать коды только тех признаков объектов, которые необходимы, что дает возможность работать в каждом отдельном случае с кодами небольшой длины.
- Можно осуществлять группировку объектов по любому сочетанию признаков.
- Метод хорошо приспособлен для машинной обработки информации.

- По конкретной кодовой комбинации легко указать, набором каких характеристик обладает рассматриваемый объект. При этом из небольшого числа признаков можно образовать большое число кодовых комбинаций.
- Набор признаков при необходимости может легко пополняться присоединением кода нового признака. Это свойство параллельного метода кодирования особенно важно при решении технико-экономических задач, состав которых часто меняется.

Параллельный метод кодирования целесообразно использовать для кодирования однородных объектов, так как в противном случае реальной становится лишь незначительная часть сочетаний признаков, и емкость классификатора будет использоваться не полностью. Это является недостатком данного метода кодирования. К недостаткам метода можно отнести также и другие недостатки, присущие фасетному методу классификации.

Перечисленные классификационные методы кодирования характеризуются тем, что даже при глубокой классификации объектов код несет информацию о классификационной группировке, но не всегда идентифицирует конкретный объект, а коды, полученные на основе регистрационных методов, хорошо выполняя функцию идентификации объектов, практически не несут информацию об их свойствах. Поэтому регистрационные и классификационные методы кодирования чаще всего применяются в классификаторах в сочетании друг с другом.

### **Контроль достоверности кодов**

Использование кодов требует обеспечения высокой степени достоверности кодированной информации. В классификаторах ТЭСИ для выявления ошибок в кодах используется метод контрольных чисел.

Контроль правильности записи кодов при обработке информации основан на принципе делимости чисел. Иначе его называют контролем по модулю. Суть метода заключается в том, что к коду добавляется еще один проверочный знак (контрольное число), связанный с кодом определенной математической зави-

симостью. При вводе кодированной информации в базу данных, ее обработке или использовании в ЭВМ специальной программой контроля выполняется проверка этой зависимости по каждому коду. Если зависимость нарушается, машина выдает информацию о наличии ошибки в коде.

Контроль по модулю широко используется в классификаторах ТЭСИ как у нас в стране, так и за рубежом. В качестве модуля используют различные числа, но наибольшее распространение получил в настоящее время контроль по модулю 11. Для общероссийских классификаторов расчет контрольных чисел осуществляется в соответствии с методикой, разработанной ВНИИКИ (Всероссийский научно-исследовательский институт классификации информации). В соответствии с этой методикой контрольным числом является остаток от деления на 11 суммы произведений весов на значения разрядов кода. Весом (весовым коэффициентом) является порядковый номер разряда в коде слева направо.

На практических занятиях мы более подробно изучим методику вычисления контрольного числа.

Методика ВНИИКИ предлагает использовать в качестве весов натуральный ряд чисел от 1 до 10. Если разрядность кода больше 10, то набор весов повторяется. При использовании данного метода остаток может получить значение от 0 до 10. Так как методика предусматривает использование одноразрядных контрольных чисел, то при получении остатка, равного 10, следует сделать повторный расчет контрольного числа со сдвигом строки весов. В этом случае весовой ряд начинается с 3 до 10, а если разрядность кода больше, то дальше веса идут с 1 до 10. В случае повторного получения контрольного числа, равного 10, в качестве контрольного числа используется 0. В случае, если сумма произведений весов на значения разрядов получается меньше 10, то эта сумма и является контрольным числом.

Использование контрольных чисел обеспечивает возможность обнаруживать и исправлять ошибки в кодированной документной информации, что повышает ее достоверность.

Для сведения: При передаче информации по каналам связи используется множество других методов контроля достоверности информации, в том числе, контроль по четности (или нечетности), вычисление контрольной суммы или сигнатуры сообщения. Кроме того, применяются специальные методы кодирования информации, позволяющие не только обнаружить ошибку, но и исправить ее (например, коды Хеминга).

Одним из наиболее узких мест во всей технологии использования классификаторов информации является кодирование и ввод данных. С целью ускорения этого проводятся исследования по автоматизации процесса кодирования информации. Однако для реализации автоматизированного процесса кодирования требуются большие объемы памяти, так как вначале вся информация вводится на естественном языке, и связанные с этим большие трудозатраты.

Другим направлением снижения трудозатрат в процессе кодирования и ускорения этого процесса является использование штриховых (линейных) кодов. Преимущества штриховых кодов состоят в следующем:

- резкое снижение числа ошибок при вводе информации в виде штриховых кодов по сравнению с вводом информации с клавиатуры;
- легкость считывания штриховых кодов электронными оптическими системами по сравнению с буквенно-цифровыми символами;
- высокая экономическая эффективность применения систем на основе штриховых кодов вследствие резкого снижения стоимости ввода данных в систему.

Штриховой (линейный) код представляет собой комбинацию вертикальных полосок разной ширины и пробелов между ними. При этом за базу принимается ширина узкого элемента (полоски) кода. Широкие полоски должны быть кратными им по ширине или находиться с ними в определенных соотношениях. В основе штрихового кода лежит цифровой код.

В разных странах используются различные виды штриховых кодов. В каждом из них установлено определенное соотношение между широкими и узки-

ми полосками и между полосками и интервалами между ними. Так, в "Коде 39" каждому знаку цифрового кода соответствует комбинация из девяти элементов (три широких полосы и шесть узких) и из них пять штрихов и четыре интервала между ними.

Разработка штриховых кодов осуществляется Европейской ассоциацией по нумерации (ЕАН) (European Article Numbering), коды которой являются наиболее распространенными в Европе. Наша страна с 1987 года также стала членом ЕАН. В 1988 году Госстандарт утвердил Руководящий документ РД 50-666-88 "Методические указания. Присвоение цифровых кодов товарам народного потребления". Этим документом устанавливались правила присвоения товарам народного потребления цифровых (торговых) кодов. Эти цифровые коды служат основой для штриховых кодов, наносимых на ярлыки, упаковку и этикетки товаров. Такой цифровой (торговый) код строится в полном соответствии с кодом ЕАН-13. Он состоит из тринадцати разрядов и имеет следующую структуру:

- 2 или 3 знака – идентификатор страны-изготовителя товара;
- 5 знаков – идентификатор фирмы-изготовителя товара;
- 5 знаков – идентификатор товара;
- 1 знак – контрольное число.

В этом коде, например, США и Канада имеют идентификаторы с 00 до 09, Франция – с 30 до 37, ФРГ – с 40 до 43, Россия и страны СНГ – 460–469, Япония – 49, Италия – с 80 до 83, Корея – 88 и так далее.

В штриховом коде, построенном на основе ЕАН-13, каждому знаку цифрового кода соответствует комбинация из семи элементов – штрихов и пробелов между ними.

/\*Штриховой (линейный) код имеет следующий вид:)





\*/

Штриховые коды могут использоваться кроме торговли также в таких областях, как медицина, банковское дело, промышленность и других. При этом в качестве цифровых кодов для них могут использоваться коды классификаторов ТЭСИ.