Использования машинного обучения для создания единого классификатора МТР

А.Н. Зыков,
инж.- прогр., alexzikov@gmail.com
АО "ЦС "Звёздочка" г. Северодвинск
С.К. Карцов,
д.т.н., проф., kartsov@yandex.ru
"Московский Политех" г. Москва

Разработка нормативно-информационных справочников как формальное описание терминов предметной области и отношений между ними стало обычным явлением в настоящее время. Справочник определяет общий словарь терминов, который использует формулировки основных понятий в предметной области и отношения между ними. В результате возникла ситуация, когда каждый такой справочник у различных организаций в одной предметной области содержит различное базовое описание одних и тех же терминов. Это усложняет задачу коммуникации между организациями в одной предметной области. Решения этой проблемы состоит в создании единого справочника, где описание терминов будет упорядочено и согласовано между собой. Это сложная задача и для её решения, предлагается использование машинное обучения, которое позволит перевести решение в автоматическую плоскость и значительно облегчить эту работу.

Development of information reference books as the formal description of terms of subject domain and the relations between them became an everyday occurrence now. The reference book defines the general dictionary of terms which uses formulation of the basic concepts in the subject domain also the relations between them. The situations when each reference book at various organizations in one subject domain various basic description and same terms resulted. It complicates a problem of communication between the organizations in one subject domain. The solution of this problem consists in creation of the uniform reference book where the description of terms will be ordered and agreed among themselves. It is a complex challenge and for its decision use machine training which allow to transfer the decision to the automatic plane and considerably to facilitate this work is offered.

Введение

В настоящее время, с использованием информационных систем на предприятиях, возникла ситуация, когда каждая такая система работает локально и никак не связанна с другими информационными системами. Каждая информационная система предприятия сегодня существует и поддерживает в основном свою систему классификаторов. Это создает проблему несогласованности между собой информационных потоков. Ярким примером несогласованности является работа различных подразделений и предприятий с классификаторами, содержащими подчас противоречивую, неполную информацию, имеющую к тому же, большое количество дублированных записей, различную базовую онтологию описания терминов предметной области. Необходимо упорядочить информацию в таких классификаторах и согласовать между собой классификаторы различных подразделений, предприятий и контрагентов, сделать это быстро и эффективно.

Как разобраться с разнообразием описания одной позиции, представленной в разных классификаторах по своему. Например, необходимо закупить определенную продукцию, мы знаем технические характеристики данной продукции и её код, они есть в нашем классификаторе, но при поиске сведений по конкретной продукции приходится перерабатывать большие объёмы информации из совершенно разнородных источников - различных ГОСТов, ОСТов, ТУ и она может иметь совершенно другое описание и код, относится к различным классификационным группам, что также требует значительных затрат времени и сил. Для удобства поиска, необходимо свести воедино информацию из различных источников и классифицировать её таким образом, чтобы каждый пользователь в ней мог легко ориентироваться и получать сведения о необходимой ему продукции - правильное название и номер ГОСТа (ОСТа, или ТУ) [1].

Единый классификатор МТР возможно использовать в системе материально-технического снабжения предприятия для стандартного обозначения продукции, например, чтобы одна и та же продукция не проходила по документам под различными названиями, и наоборот, одним и тем же названием не обозначалась различная номенклатура.

1. Понятие онтологии

В сложившейся ситуации очень кстати приходится термин - "онтология". Онтологии широко используются во всех областях, занимающихся обработкой данных на естественном языке. Под онтологией понимается система понятий некоторой предметной области, которая представляется как набор сущностей, соединенных различными отношениями. Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную предметную область знаний. Преимуществом онтологий в качестве способа представления знаний является их формальная структура, которая упрощает их компьютерную обработку [4].

В общем виде структура онтологии представляет собой набор элементов следующих категорий:

- классы (понятия)
- отношения (свойства, атрибуты)
- функции
- аксиомы
- экземпляры

Классы или понятия используются в широком смысле. Понятием может быть любая сущность, о которой может быть дана какая-либо информация. Классы - это абстрактные группы, коллекции или наборы объектов. Они могут включать в себя экземпляры, другие классы, либо же сочетания и того, и другого. Классы в онтологиях обычно организованы в таксономию - иерархическую классификацию понятий по отношению включения.

Отношения представляют тип взаимодействия между понятиями предметной области.

Функции — это специальный случай отношений, в которых *n*-й элемент отношения однозначно определяется *n*-1 предшествующими элементами.

Аксиомы используются, чтобы записать высказывания, которые всегда истинны. Аксиомы задают условия соотнесения категорий и отношений, они выражают очевидные утверждения, связывающие понятия и отношения.

Экземпляры - это отдельные представители класса сущностей или явлений, то есть конкретные элементы какой-либо категории [4].

Онтологии сильно различаются по ряду параметров в зависимости от набора элементов, содержащихся в них, а также типов вводимых отношений и исследователи выделяют различные основания для их классификации.

Для создания онтологии надо сначала перечислить категории, обозначающие сущности или явления в моделируемой области. Затем следует связать эти категории определенными отношениями. И на последнем шаге надо соотнести категориям набор конкретных экземпляров.

При создании онтологии, необходимо также определить, какому дифференцированному признаку отдать предпочтение при классификации, создавать новый элемент и должен ли этот элемент быть включен в структуру. Следует понять, где по отношению к другим объектам должен располагаться вновь создаваемый элемент, должен ли он быть видом какого-либо класса или же сам представляет собой корневым понятием. Помощь в этом может оказать формулировка особых, уникальных свойств элемента, то есть его отличительных характеристик. При этом не следует смешивать свойства понятия и его отличительные признаки.

Онтология представляет собой дерево, где на верхнем уровне описаны самые общие понятия, что такое пространство или что такое время, а где-то ниже идет специализация по областям знаний и предметным областям, например, будет написано "насос", а где-то там дальше написан "насос на конкретном предприятии".

Для примера, возьмем какой-либо набор терминов в предметной области. Как они связаны с реальной жизнью? Ни одной привязки терминов к реальной жизни нет. Чтобы это понять, необходимо откинуть значения слов, посмотреть на их определения и попытаться переформулировать и тогда выясниться, что термин из одного набора данных относится туда, куда и термин из другого набора данных. И другие термины из разных наборов данных также соотносятся друг с другом (рис. 1). И мы говорим, что вот эти пары терминов имеют одни и тоже понятия, и мы называем их какими-либо словами.

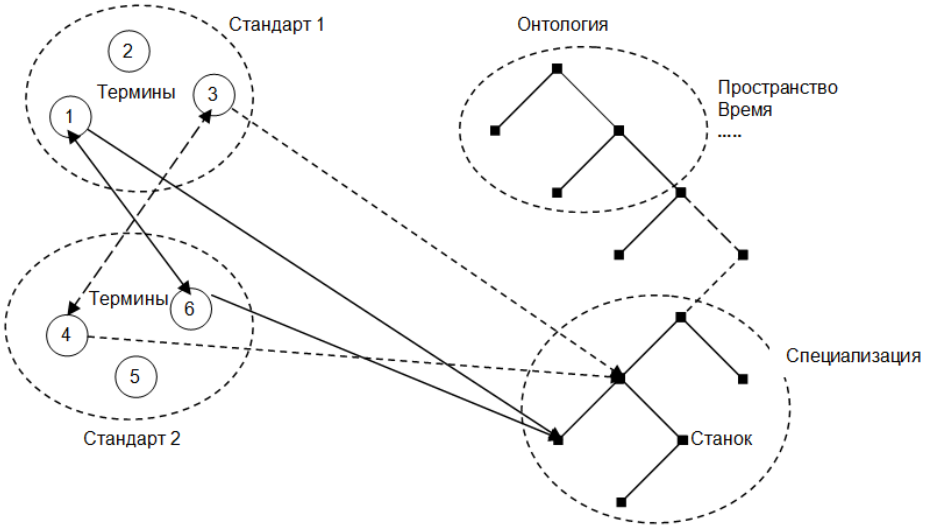

рис. 1 Схема привязки терминов к понятиям реальной жизни

Для создания единой онтологии из нескольких источников необходимо провести определенные шаги над имеющейся информацией. Например, соединение двух онтологий происходит в несколько этапов. На начальной стадии необходимо найти связующее звено, на основании которого возможно произвести слияние. Затем выравнивающий алгоритм по описанию элементов находит их место в новой структуре. Потом происходит выверка результатов слияния в контексте.

Выделяют несколько различных методов обнаружения связующих звеньев.

- текстовые совпадения - идентичность имен понятий, текстовых определений.
- совпадение иерархических отношений - поиск общих вышестоящих понятий, нахождение семантического расстояния.
- совпадение форматов и данных опирается на отношения внутри понятий.

После отработки алгоритмов слияния, проверяются результаты работы всех процедур, и выдается общий коэффициент совпадения. Для выявления идентичности понятий используются специально созданные критерии - материал, форма, части из которых сделан экземпляр, его функциональное использование [2].

Слияние онтологий может также использоваться для устранения проблемы дублирования и возможности выявить ошибки и опущения в онтологиях.

Классификационная структура (таксономия) является неотъемлемой частью любой онтологии и можно говорить о присутствии элементов онтологий в любых классификациях. Онтология обеспечивает непротиворечивое накопление любого количества информации в стандартной структуре классификации.

2. Определение классификатора

Классификатор по своему типу является систематизированным перечнем наименованных объектов, каждому из которых в соответствие дан уникальный код. Классификация объектов производится согласно правилам распределения объектов на классификационные группы в соответствии с установленными признаками их различия или сходства. Классификатор является средством единообразного обозначения товаров и материалов, для всех заинтересованных сторон. Правильно разработанный классификатор обладает следующими достоинствами:

- Простота использования при оформлении документации.
- Распределение наименований товаров на основе общности свойств.
- Созданию возможности для снижения уровня запасов товаров на предприятии и их рационального использования.

При разработке нового классификатора неотъемлемым результатом процедуры классификации изделия должно явиться соответствие этого изделия конкретному утвержденному наименованию из единого перечня стандартных наименований продукции (ЕПСН). ЕПСН или "Утвержденное наименование" (УН) является важнейшим понятием в системе классификации продукции, а формирование и использование единого перечня УН - важнейшим этапом в процессе классификации изделий.

Необходимо также определиться на базе, каких общероссийских классификаторов (ОК) целесообразнее разрабатывать новый классификатор МТР. Тут есть несколько вариантов [1]:

- Построение корпоративного классификатора продукции на основе ОКПД2/ОКПД/ОКП/ОКДП/ОКОФ.
- Построение корпоративного классификатора продукции на основе ЕКПС для государственных закупок в системе ФСКП.
- Построение корпоративного классификатора продукции на базе классификаторов стандартов ОКС/МКС.
- Построение корпоративного классификатора продукции на базе классификатора международной системы КОМПАСС.

В этих классификаторах используется различная, иерархическая классификация с различной цифровой системой кодирования. Кодирование является присвоением по определённым правилам изделию или группе изделий кодов. Это позволяет заменить полное наименование этих изделий несколькими знаками и использовать коды для однозначной идентификации объектов учёта, представить информацию в удобной для сбора и передачи форме, приспособить ее к автоматизированной обработке, а также обеспечить поиск, сортировку и группировку конкретных данных.

Применение вместо названия продукции обозначающих её классификационных кодов позволяет производить в масштабах предприятия отслеживание всего жизненного цикла потребности в МТР от момента сбора заявок до момента списания в производство. Таким образом, совокупность правил и методов классификации и кодирования находят совместное применение в классификаторах.

Структура классификатора включает позицию и емкость. Позиция классификатора – это наименование и код классификационной группировки. Емкость - наибольшее число позиций, которое может содержать классификатор. Например, в ОКПД2 (ОК 034-2014) использованы иерархический метод классификации и последовательный метод кодирования с цифровой десятичной системой кодирования. На каждой ступени классификации деление осуществляется по наиболее значимым экономическим и техническим классификационным признакам. Например, количество актуальных записей в базе ОКПД2 составляет 19 113 (таблица 1).

Таблица 1

Ёмкость ОКДП2 по уровням иерархии

Разрядность кода	Уровень иерархии	Кол-во группировок
XX	Класс	88
XX.X	Подкласс	273
XX.XX	Группа	618
XX.XX.X	Подгруппа	1464
XX.XX.XX	Вид	3215
XX.XX.XX.XX	Категория	7786
XX.XX.XX.XXX	Подкатегория	5914

Число разделов в ОКПД2 составляет 21 и литеры обозначения разделов (с A до U) в кодировании продукции не участвуют. Код идентификации продукции может включать от двух до девяти цифр с разделителем «точка» между структурными единицами классификатора (рис. 2).

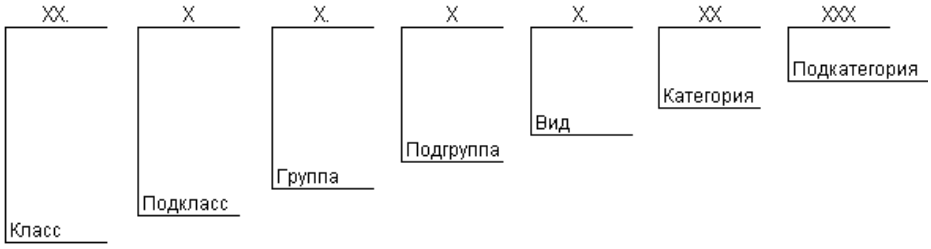

рис. 2 Структура классификатора ОКПД2

Например, раздел "A" включает:

- Использование растительных и животных природных ресурсов, включая выращивание зерновых, содержание и разведение животных.
- Получение древесины и других растений, животных или продуктов животного происхождения на ферме или в естественной среде обитания.

Класс 02 в ОКПД2 означает классификационную группировку «Продукция лесоводства, лесозаготовок и связанные с этим услуги» и так далее по иерархии.

После выбора исходного ОК необходимо выполнить комплекс работ по созданию перечня базовых классов МТР, которые будут являться основой для последующего кодирования продукции.

3. Моделирование алгоритма

Для создания нового классификатора предлагается использование машинного обучения. Машинное обучение включает множество различных подходов и алгоритмов. Смысл машинного обучения состоит в использовании выбранных признаков для построения моделей, подходящих для решения поставленных задач. Признак определяет язык, на котором описываются объекты предметной области. Имея представление в виде признаков, мы уже можем не возвращаться к самим объектам предметной области. Задачу машинного обучения можно представить как отображение исходных данных на результаты [5]. Это отображение или модель является результатом работы алгоритма машинного обучения, применённого к обучающим данным (рис. 3).

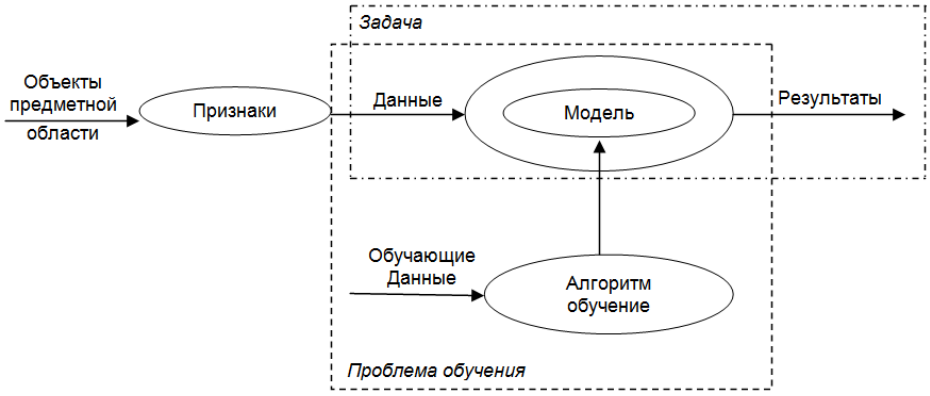

рис. 3 Применение машинного обучения

Исходя из всего выше сказанного, приведем простой алгоритм:

1. Обучить нашу модель с помощью выбранного классификатора.
2. Нормализовать данные в выбранном источнике.
3. Подать эти данные на вход модели и получить классификационную метку.
4. Соединить в единую структуру и устранить дубликаты.

Как мы видим, первым шагом необходимо провести обучение нашей модели на актуальных данных по структуре классификатора. Вторым шагом проведение нормализации информации по каждой позиции в выбранном справочнике. Проблема состоит в том, что каждый справочник имеет свою структуру полей и их заполнение. Поэтому первым шагом необходимо выбрать нужные признаки для построения модели. Так как признаки описывают объекты предметной области, а область у нас некий справочник, то признаками будут атрибуты справочника, например, код, наименование, обозначение, номенклатурный номер, группа и подгруппа и даже примечание. На основании этих признаков нам надо так настроить и обучить нашу модель, чтобы она более точно выдавала на выходе значение, которое наилучшим образом будет приближать позицию справочника к определённой позиции в нашем будущем классификаторе.

Ф45 08Х18Н10Т-ВД ГОСТ2590-88
ПЛИТКИ 150Х50 ЧЕРНАЯ ТИП24 ГОСТ6141-91
ПРОКЛАДКА Д102ХД92,5Х0,2 БУМАГА ЧЕРТЕЖНАЯ
521-35.3512-02
СОЕДИНЕНИЕ ТРУБНОЕ

рис. 4 Пример описания позиции данных

Представленные на рис. 4 примеры, показывают, как могут быть описаны определённые позиции объектом в базе данных предприятия. Описание объекта может быть определено в одном или разбито по нескольким атрибутам. Какие здесь выделить признаки для решения задачи классификации? Первое что приходит на ум, мы можем, например, разбить данные конструкции на токены - части слов каким-либо способом, для дальнейшей передачи их в качестве признаков в модель машинного обучения. По некоторым позициям почти всё ясно, указан гост, название, свойства объекта, а по другим данных совсем немного.

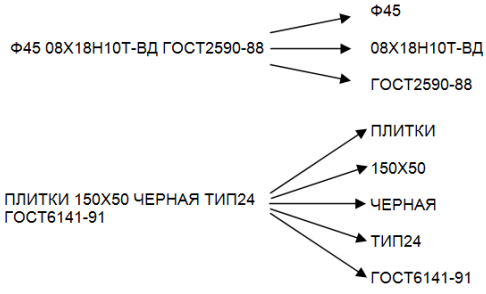

рис. 5 Разложение обозначения объекта на токены

Следующим шагом будет определение, к чему относятся данные токены - обозначению, наименованию, характеристике, виду стандарта и т.д. (рис. 6). После определения этого можно подать на вход модели полученные признаки для каждой позиции и получить на выходе классификационную метку.

Код группы		Наименование группы				
Ед. изм.	Код под-группы	Наименование материала	№ ГОСТ Или ТУ на сортамент	Характеристики	№ ГОСТ Или ТУ на марку	Код типоразмера
1	2	3	4	5	6	7

рис. 6 Отраслевой классификатор материалов

Для построения модели, в качестве основного алгоритма можно выбрать несколько подходов и использовать различные алгоритмы машинного обучения. Выбор алгоритма исходит не с точки зрения быстродействия или потребления памяти, а правильности классификации.

Решение этой задача относится к решению задаче многоклассовой классификации, так как у нас количество классов равно ёмкости нашего классификатора. Согласно [5] для этой цели лучше всего подходит метод деревьев признаков, метод *k* - ближайших соседей (*k*NN) и наивный байесовский классификатор. Но по тому, как модель может быть группирующей, приспособлена к обработке дискретных данных, многоклассовой классификации, обучению с учителем и без, то выбор падает на модель деревьев признаков. Те деревья, у которых листья помечены классами, как в нашем случае, называются решающими деревьями. Принцип их работы - разбиение пространства объектов на все более мелкие подмножества. Все объекты, попавшие в один листовой узел, трактуются одинаково, независимо от неотраженных в дереве признаков, для их дальнейшего различия. В некоторых случаях имеет смысл использовать лес или джунгли решающих деревьев.

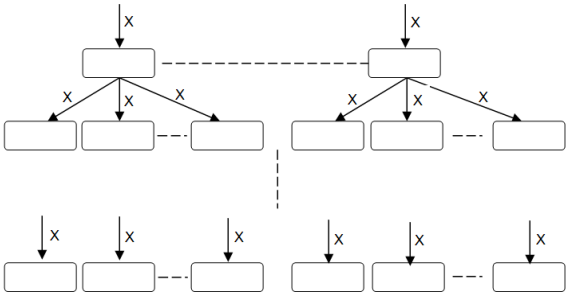

рис. 7 Решающие деревья

При работе данной модели можно сразу провести кодификацию объекта. Для этого каждый узел определённого уровня имеет свою позицию и номер кода в иерархии (рис. 7). Если на каком-то уровне не хватает признаков для перехода на более низкий уровень, то останавливаемся на нём и записываем данный объект на место в классификаторе, которое обозначает его полученный код. В итоге мы получаем иерархическую структуру справочника, теперь уже классификатора и нам осталось убрать только дублирующие записи.

Стоит признать, что машинное обучение, конечно, не может полностью произвести классификацию данных из предложенных источников. Часть работы все равно остается за оператором. Он должен в ручную классифицировать некоторое множество данных, в виду того, что признаки, используемые для описания данных, дают лишь предположительное указание о возможном классе, но не дают сильного сигнала, который позволил бы с уверенностью предсказать класс и автоматически провести классификацию. Работа оператора также позволяет более точно настроить алгоритм работы модели и тем самым провести более точное обобщение его работы на предложенных данных. Стоит сказать, что данная задача очень важна и актуальна в настоящее время и повышение точности работы построенной модели является очевидным фактом для продолжения работы в данном направлении.

Заключение

Разработка единого классификатора МТР лучше послужит для общего понимания структуры информации в предметной области пользователями системы. Единый классификатор МТР необходимо использовать в системе материально-технического снабжения предприятия для стандартного обозначения продукции, например, чтобы одна и та же продукция не проходила по документам под различными названиями, и наоборот, одним и тем же названием не обозначалась различная номенклатура. Применение вместо названия продукции обозначающих её классификационных кодов позволяет производить в масштабах предприятия отслеживание всего жизненного цикла потребности в МТР от момента сбора заявок до момента списания в производство, и машинное обучение имеет возможность перевести большую часть работы по созданию единого классификатора МТР в автоматическую плоскость.

Литература

1. Концепция создания системы централизованного ведения единого номенклатурного справочника и единого классификатора МТР / ФГУ «Федеральный центр каталогизации», Москва, 2006
2. Боргест Н. М., Онтология проектирования / Самара, 2010
3. Федеральная система каталогизации продукции для федеральных государственных нужд / ГОСТ Р 51725.1-2012
4. Константинова И.С., Митрофанова О.А.. Онтологии как системы хранения знаний / СПГУ Факультет филологии и искусств, Кафедра математической лингвистики, С.-Петербург
5. Флах П., Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Москва, ДМК Пресс, 2015