УДК 004.65 ББК 32.97

Е.Е. Раковская*

ПРИМЕНЕНИЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ ПОСТРОЕНИИ БАЗ ДАННЫХ ОБОРУДОВАНИЯ

В условиях постоянного увеличения объема информации критически важным становится качество данных. Наличие в базах данных «неполной» или «некорректной» информации значительно усложнило функционирование информационных систем, что, в свою очередь, вызвало необходимость эффективной очистки информации. Для решения задачи извлечения полезных сведений из баз данных перспективными являются методы искусственного интеллекта. В статье рассматриваются некоторые аспекты применения методов искусственного интеллекта для обработки и фильтрации информации баз данных оборудования.

Ключевые слова: искусственный интеллект, обработка информации, алгоритм Левенштейна, критерий схожести

Искусственный интеллект – научное направление, связанное с моделированием человеческих интеллектуальных функций. В настоящее время «искусственный интеллект» – ветвь информатики, имеющая как фундаментальные, чисто научные основы, так и развитые технические, прикладные аспекты, связанные с созданием и эксплуатацией работоспособных образцов информационных систем. Любая задача, где не известен алгоритм решения, может быть отнесен к сфере искусственного интеллекта. Особенности задач искусственного интеллекта – преобладающее использование информации в символьной (а не в числовой) форме и наличие выбора между многими вариантами в условиях неопределенности.

Одно из направлений, где применяются методы искусственного интеллекта – это анализ и генерация текстовой информации, ее понимание, выявление значений. Трудности применения связаны, в данном случае, с тем, что часть текстовой информации не выражается определенно и ясно. Информации присуща неполнота, неточность, нечеткость, грамматическая некорректность, избыточность, зависимость от контекста, неоднозначность. В технических системах должен использоваться формальный язык, смысл значений которого однозначно определяется их формой. Таким образом, предварительная обработка текстовой информации является необходимым условием для ее дальнейшего применения.

Очистка от шумов и сглаживание рядов данных. Часто ряды данных содержат случайные изменения значений, которые можно рассматривать как шум. Шум мешает выполнять выборку данных, делает неустойчивой работу алгоритмов, не позволяет обнаруживать в данных скрытые закономерности,

* Раковская Елена Евгеньевна – аспирант, кафедра информатики и кибернетики, Байкальский государственный университет экономики и права, г. Иркутск, rakovskaya19@mail.ru.

структуры, тенденции и т.п. Сглаживание необходимо в том случае, когда ряд данных оказывается неравномерным, содержит большое количество мелких структур (которые не являются шумом в обычном понимании).

Например, для термопреобразователей сопротивления применяются следующие обозначения: ТСМ, ТСП, ТСПУ, ТСМУ, КТПТР. При анализе введенных значений могут быть выявлены следующие ошибки: наличие прописных и строчных букв (Тсм, ТСп, ТСПу), имеются опечатки (ТСМю, ТСМ9). Чтобы выявить эти ошибки, поля ввода проверяются на наличие строчных букв, запятых, цифровых символов, аномальных сочетаний. В зависимости от найденных ошибок строчные буквы в названиях заменяются соответствующими прописными буквами («Тсм» на «ТСМ»), очищаются от лишних знаков («ТСМю» на «ТСМ», «ТСМ9» на «ТСМ»). Здесь были применены формальные правила для поля «Обозначение термообразователей сопротивления», которые можно записать так: «Если поле состоит не только из прописных букв, то лишние символы удаляются, или строчные буквы преобразуются в прописные».

Восстановление пропущенных записей. Пустые значения вызывают неопределенность при работе многих алгоритмов. Даже одно пропущенное значение может привести к невозможности применения программных средств. Если же пропущенных данных много, объема информации в выборке может оказаться недостаточно.

Редактирование аномальных значений. Аномальные значения также требуют внимания при подготовке данных. В большинстве случаев они являются просто ошибками ввода. Ответы на вопросы, какие значения считать аномальными, принимать ли меры к их подавлению, каковы должны быть степень и методика подавления, далеко не однозначны и требуют дополнительных исследований в контексте заданных условий.

Снижение размерности входных данных. В работе большинства моделей лежит принцип обобщения. То есть, чтобы получить на выходе даже единственное значение, нужно подать на ее вход некоторый набор данных, на основе соотношений между которыми и будет определено выходное значение. При создании баз данных изначально привлекается максимум собранной информации об исследуемом объекте, в результате набор входных переменных разрастается, что приводит к усложнению модели. Поэтому одной из важных задач подготовки данных является снижение их размерности.

Обработка дубликатов и противоречий. Неуклонный рост объемов данных вызывает необходимость широкого использования передовых информационных технологий для эффективного управления потоками данных. При этом наибольшую значимость приобретают задачи создания эффективных инструментов оценки и контроля растущих потоков информации, оптимизации процедур обработки, агрегации, обобщения, поиска и анализа данных. Существуют проблемы управления качеством данных, т.е. обеспечение такого состояния информации, которое удовлетворяет требованиям пользователя по критериям достоверности, актуальности, логической полноты и непротиворечивости, отсутствие дублирующей информации.

Для выполнения всех перечисленных требований понадобится целый комплекс мер, в настоящее время универсальной методики не существует, поскольку каждая проблема имеет свою специфику. Вследствие этого задача текстовой идентификации в базах данных не может быть в полной мере решена только методом проверки на точное соответствие. Становится актуальной задача разработки специальных методов и технологий поиска с использованием нетривиальных решений, в т.ч. с использованием операции нестрогого соответствия. Одним из применений метода неполного соответствия является обеспечение отсутствия дублирующей информации. Проблема дублирования часто решается применением словарей-справочников, если их требуемый объем имеет разумные пределы и не превышает объем основной информации. Для проверки дублирования информации (сюда входят данные, имеющие идентичный набор значений всех признаков, синонимические записи и др.) применяются алгоритмы нечеткого поиска, позволяющие находить данные на основании неполного совпадения и оценки их релевантности – количественного критерия схожести.

Большинство количественных методов основаны на анализе лингвостатистических (числовых) характеристик, вычисляемых по тексту.

Можно выделить следующие характеристики, которые могут быть получены для технических текстов и доступны для последующего анализа:

1. Количество (частота) появления в названиях единиц оборудования строчных и прописных букв, также частота появления знаков, символов, цифр.

2. Однородность текста, или, распределение составляющих единиц текста – наличие буквенных сокращений, чисел, полных названий.

3. Анализ дополнительных признаков текста – аббревиатур, знаков пунктуации, символов.

Полученные лингвостатистические характеристики подвергаются анализу с использованием различных математических методов, среди которых можно выделить следующие: статистические методы, методы распознавания образов и искусственного интеллекта.

Количественным критерием схожести является расстояние между строками – наименьшее число операций, модификаций над буквами для преобразования одной строки в другую. Чем меньше проделано операций по преобразованию одного слова в другое, тем больше вероятность того, что найдено именно нужное слово.

В алгоритме Левенштейна основная идея заключается в том, чтобы рассчитать минимальное количество операций удаления, вставки, и замены, предназначенных для преобразования одной строки в другую.

Например, термометры сопротивления, в конструкции которых применены полупроводниковые материалы, называют термосопротивление, термистор, терморезистор. Верное написание «терморезистор» и неверное – «терморезистр». Найдем расстояние между ними, подсчитав количество необходимых операций для преобразования одного в другое.

«Терморезистр» – «Терморезистор» (операция вставки).

Таким образом, расстояние между названиями равно единице.

На этом же примере рассчитаем максимальную общую последовательность двух строк (получаем 12). Чем больше максимальная последовательность двух слов (12 из 13 возможных), тем больше вероятность того, что это одно и то же слово.

Общий принцип применения алгоритмов для поиска дубликатов следующий.

1. Производится вычисление некоторого показателя соответствия («похожести») двух символьных строк, например, дистанции (расстояния) или релевантности.

2. Данный показатель относится к шкале соответствия в интервале от 0 до 1 (0 – полное несовпадение, 1 – полное совпадение). Эта шкала может быть приведена к процентному виду, удобному для восприятия человеком.

3. Для набора данных определяется нижний порог автоматической обработки, за которым количество ошибок распознавания дубликатов становится неприемлемым, т.е. за которым поиск выдает практически одни ошибки.

Задачу выявления и устранения дубликатов можно разбить на этапы:

1. Выявление дубликатов на уровне ввода информации пользователями и их отклонение.

2. Выявление дубликатов путем сравнения и анализа уже введенных данных в соответствии с заданным порогом автоматической обработки и удаление дублирующей информации.

3. Анализ и обработка человеком результатов предыдущего этапа, которые не могут быть обработаны автоматически.

Практическое применение методов искусственного интеллекта заключается в создании процедур и функций, а также комплекса алгоритмов поиска и сравнения записей в базах данных, которые позволяют:

1. Осуществить расширенный поиск и выдачу информации на основе функций нестрогого соответствия;

2. Идентифицировать записи в БД, содержащих информацию о типах, марках оборудования, информацию об изготовителях и поставщиках;

3. Проводить быструю оценку, обобщение и агрегацию данных, обеспечить возможность интеллектуального анализа;

4. Повысить уровень информационного обеспечения пользователей за счет снижения зашумленности данных.

Автор благодарит проф. А.В. Боровского за постановку проблемы и консультации.

Список использованной литературы

1. Беллман Р. Вопросы анализа и процедуры принятия решений: пер с англ. / Р. Беллман, Л. Заде. – М. : Мир, 1976. – 240 с.

2. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям / Н. Б. Паклин, В. И. Орешков – М. : Вильямс, 2013. – 701 с.

3. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация: пер. с англ. / Э. Спирли. – М. : Питер, 2001. – 400 с.