УДК 004.023; 519.85

КЛАССИФИКАЦИЯ ДАННЫХ ПО СТЕПЕНИ ИХ ОТКРЫТОСТИ НА ПРИМЕРЕ ЭЛЕКТРОННОГО ДЕКАНАТА

Иванова М.И., Павленко А. А.
к.т.н., доцент каф. СПО Блажко О. А.
Одесский национальный политехнический университет, УКРАИНА

АННОТАЦИЯ. В данной работе рассматривается процесс автоматизации метода классификации информации о студентах по степени ее открытости. В качестве классификатора использована нейронная сеть. Для примера описан процесс определения степени открытости данных в электронном деканате Института компьютерных систем Одесского национального политехнического университета.

Введение. Электронный документооборот в настоящее время является неотъемлемым элементом информационной структуры любой организации. Институт компьютерных систем использует следующие электронные ресурсы.

Подсистема «Электронный Деканат» - представляет собой программный продукт, который дает возможность полностью автоматизировать процессы создания документации в каждом деканате учебного заведения.

Система контроля выставленных баллов – ресурс, позволяющий студентам просматривать выставленные баллы, а также оповещать сотрудника деканата о найденных несоответствиях.

Портал dl.opu.ua – это информационный ресурс взаимодействия студента и преподавателя, на котором преподаватели регистрируют новый курс, размещают обучающий материал, а затем студенты, имея соответствующие права доступа, могут записаться на курс и работать с ним: изучать информационные материалы, участвовать в семинарах, обсуждениях, проходить тестирование, сдавать экзамены и т.д.

E-learning - система, позволяющая контролировать актуальность информации, а также обучение студентов в онлайн-режиме. В Интернет ежедневно выставляется большое количество информации, которая нуждается в подтверждении ее достоверности. В многопользовательской системе можно использовать систему рекомендаций, которые представляют связь между двумя компонентами информации. Для предотвращения случаев предоставления пользователем сознательно неточных рекомендаций, другие пользователи могут оставлять жалобы на рекомендации, чтобы обратить внимание модераторов на некорректные рекомендации и тех пользователей, которые их предоставили.

Так как в случаях с E-learning, dl.opu.ua и electronic register данные предоставляются в открытом доступе, необходимо для начала определить степень конфиденциальности, а потом уже выставлять данные в электронных ресурсах.

Целью работы является классификация данных по степени их открытости в соответствии с законом Украины [1] на примере электронного деканата Института компьютерных систем Одесского национального политехнического университета.

Основная часть работы.

Для достижения данной цели необходимо выполнить следующие задачи:

– апробация входных данных;

– реализация классификатора.

Для начала рассмотрим содержимое БД четырех информационных систем. В БД электронного деканата хранится наиболее полная информация о студентах, а именно: - фамилия, имя, отчество;- дата, место рождения;- сведения о гражданстве; - паспортные данные;- адрес регистрации;- номер телефона;- почтовый ящик;- номер телефона родителей;- сведения о социальных льготах;- сведения о форме обучения;- сведения об основе обучения. Информационный ресурс «Электронная ведомость успеваемости» содержит следующую персональную информацию:- фамилия, имя, отчество;- логин, пароль;- группа; -

специальность; - почтовый ящик; - сведения об основе обучения; - сведения о полученных баллах. В базе данных E-learning хранится: - фамилия, имя, отчество; - группа. Для функционирования dl.opu.ua необходима следующая информация о студентах: - фамилия, имя, отчество; - группа; - специальность; - полученные баллы.

В соответствии с Законом Украины следующие персональные данные можно отнести к открытым: фамилия, имя, отчество, дата, место рождения, сведения о гражданстве, сведения о социальных льготах, сведения о форме обучения, сведения об основе обучения, остальные данные являются конфиденциальными. Также следует учитывать то, что если на вход мы получаем конфиденциальную информацию с открытыми данными, на выходе степень открытости соответствует данным второго типа.

Для реализации классификатора использована нейронная сеть. Как правило, нейронные сети оказываются наиболее эффективным способом классификации, потому что генерируют фактически большое число регрессионных моделей [2-3].

Чтобы информацию можно было подать на вход нейронной сети, необходимо решить несколько задач.

Во-первых, необходимо представить все слова в виде цифр. Эту задачу решим с помощью метода bag-of-words - кодируем вектором длины N (размер словаря), для каждого примера i-й элемент вектора равен количеству этих слов\символов в тексте.

Так как векторы должны быть одинаковой длины, для решения поставленной задачи выбираем размер словаря равный 40. Где каждый номер вектора кодирует соответствующий символ: буквы алфавита, цифры, «+», «-», « », «.».

Для обучения нейронной сети была создана обучающая выборка из 5955 примеров. Для фамилии, имени, отчества, электронной почты, место рождения используем словари и существующие базы данных, Для кодирования паспортных данных, сведений о форме и основе обучения, баллах – программно создаем выборку.

Используемая нейронная сеть состоит из 40 входных, 20 скрытых и 1 выходным нейронов.

Количество эпох – 1000.

Тип модели нейронной сети- многослойный персептрон.

Для обучения используем метод обратного распространения ошибок. В качестве симулятора была выбрана программная среда *STATISTICA Neural Networks*.

Результат работы нейронной сети, выбранной архитектуры, показал среднеквадратическую ошибку, равной 0,0091.

Выводы. В данной работе описан метод классификации персональных данных студентов по степени их открытости с учетом решения следующих задач:

– сформированы требования к предварительной обработки данных с учетом Закона Украины;

– собрана обучающая выборка для нейронной сети;

– разработана и протестирована нейронная сеть классификатора.

Учитывая то, что в результате проверки сети среднеквадратическая ошибка составила 0.009, можно сделать вывод, что разработана качественная модель нейронной сети для решения поставленных задач.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. ЗАКОН УКРАЇНИ «Про захист персональних даних» [Электронный ресурс]. – Режим доступа: http://zakon2.rada.gov.ua/laws/show/2297-17

2. С. Хайкин, «Нейронные сети: Полный курс». Издатель: Издательство "Вильямс", 2008. – 1103 с.

3. Многослойные персептроны. [Электронный ресурс]. – Режим доступа: URL: http://neuronus.com/theory/240-algoritmy-obucheniya-iskusstvennykh-nejronnykh-setej. Название с экрана.