А.В. Астафьев, А.А. Демидов, У.А. Демидова
Муромский институт (филиал) федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, Владимирская область, г. Муром, ул. Орловская, д.23
E-mail: Alexandr.Astafiev@mail.ru

Разработка алгоритма идентификации складируемых промышленных изделий на основе RFID-идентификации для построения систем автоматической инвентаризации

Одним из важнейших элементов инвентаризации является процесс идентификации хранимых изделий. Однако эффективность работы современных методов и подходов к идентификации хранимых изделий очень зависят от построения технологического процесса предприятия. Исходя их этого, решение научно-технических задач, состоящих в разработке новых и совершенствовании существующих методов и средств идентификации продукции с целью повышения эффективности работы систем автоматической инвентаризации является актуальной научно-технической задачей.

Цель работы: Разработка алгоритма идентификации складируемых промышленных изделий на основе RFID-идентификации для построения систем автоматической инвентаризации.

Разрабатываемый алгоритм идентификации складируемых промышленных изделий можно условно разделить на три этапа:

1. Алгоритм подключения к RFID-считывателю;
2. Алгоритм формирования запроса на считывание;
3. Алгоритм обработки полученного результата.

1. Алгоритм подключения к RFID-считывателю

При подключении к PR9200 (рисунок 1) не обходимо знать параметры, при которых устройство будет обмениваться данными с персональным компьютером. Основными параметрами для подключения являются скорость BaudRate = 38400, а также имя COM-порта PortName (например COM3) к которому подключен считыватель.

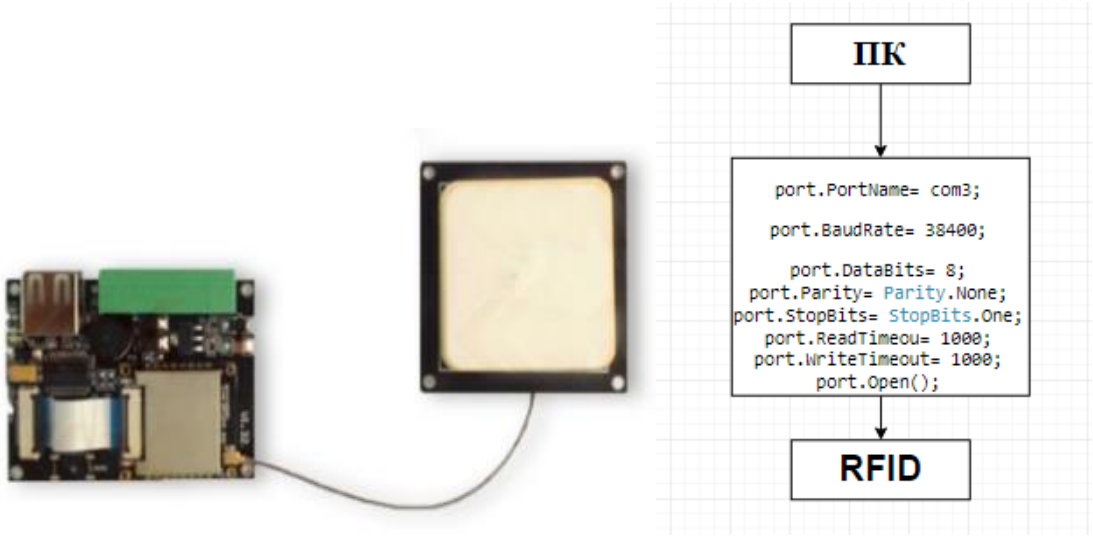

Рисунок 1 – RFID считыватель с антенной и параметры для подключения

2. Алгоритм формирования запроса на считывание

Для получения данных со считывателя мы формируем необходимую команду, состоящую из цифр в массив и передаем на устройство. После получение верной команды считыватель считывает все радиочастотные идентификаторы в своем радиусе и отправляет данные о них на ПК. В случае неверной команды считыватель остается в состояние покоя, до тех пор, пока не распознает верную комбинацию чисел, входящих в получаемый им массив. (Рисунок 2).

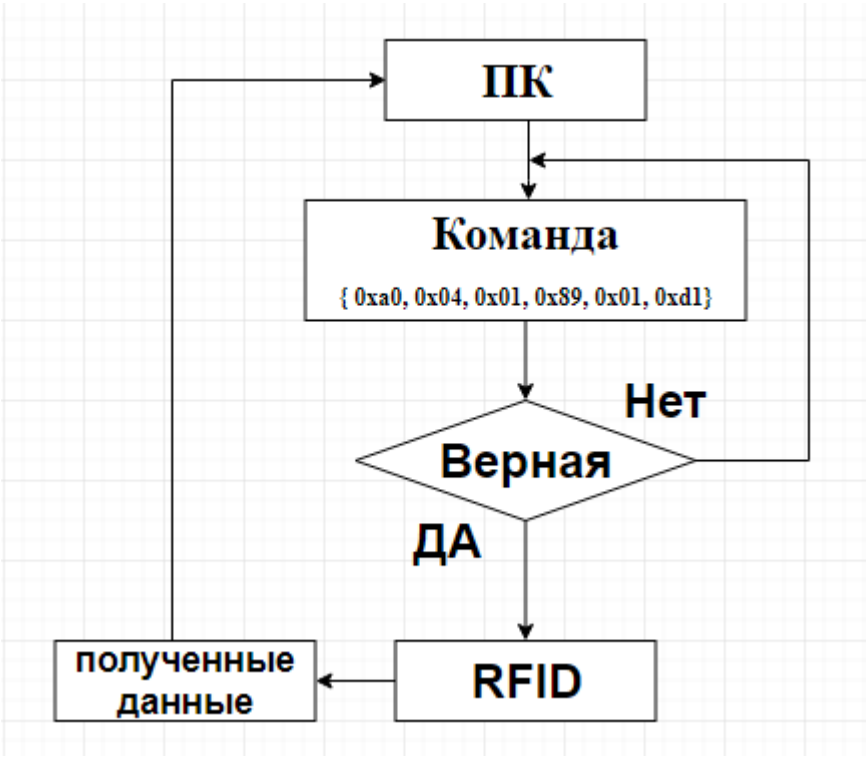

Рисунок 2 – Метод обмена данных.

3. Алгоритмы обработки полученного результата

Данные полученные от считывателя заносятся в массив. Пакет данных, содержащий информацию об одной метке имеет размер в 33 символа и имеет следующую структуру: 1 - 6 – служебная информация, 7 - 19 – UID метки, с 20 - 23 показатель уровня минимального принимаемого сигнала, 24 - 27 показатель уровня максимального принимаемого сигнала, 28 - 33 служебная информация.

Пример получаемой метки:

a0 13 01 89 60 30 00 12 12 12 12 12 10 02 33 11 ..‰`0.....3.

00 a9 86 45 0f a0 0a 01 89 00 00 10 00 00 00 01 bb .©†E. ..‰.....

В случае, если считыватель обнаружил несколько идентификаторов, они записываются последовательно (рисунок 3) и размер такого пакета S можно определить по формуле:

$$S = 33 * n,$$

где n – количество распознанных меток.

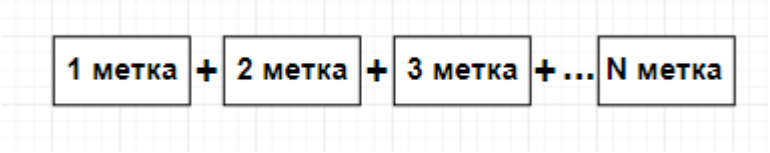

Рисунок 3 –Метод записи нескольких меток.

Помимо информации об обнаруженных метках считыватель может возвращать служебную информацию. Размер такого пакета не превышает 10 символов.

Для получения данных об обнаруженных метках предлагается использовать следующий алгоритм:

1. Подключаем считыватель к ПК;
2. Формируем команду и отправляем на считыватель;
3. Считыватель проверяет команду, если правильная начинает считывать радиочастотные идентификаторы, если не правильная остается в состояние покоя;
4. Формируется пакет данных;
5. Проверяем количество символов в пакете данных, если больше 30 – метка есть, в случае меньше 30 – метки нет;

6. Проверяем на количество меток в пакете данных (если больше 34 символов – несколько меток, в случае меньше 34 – метка одна);
7. При определении одной метки выделяются с 7 по 19 символы (UID метки);
8. Если определяется несколько меток, то выделяется с 7 по 19 символы с последующей прибавкой шага +32 (т.е. с 7+32 по 19+32, 7+32+32 по 19+32+32 и т.д.) до тех пор, пока не найдем пустое значение. Пустое значение означает что меток более не обнаружено;
9. Вывод на ПК

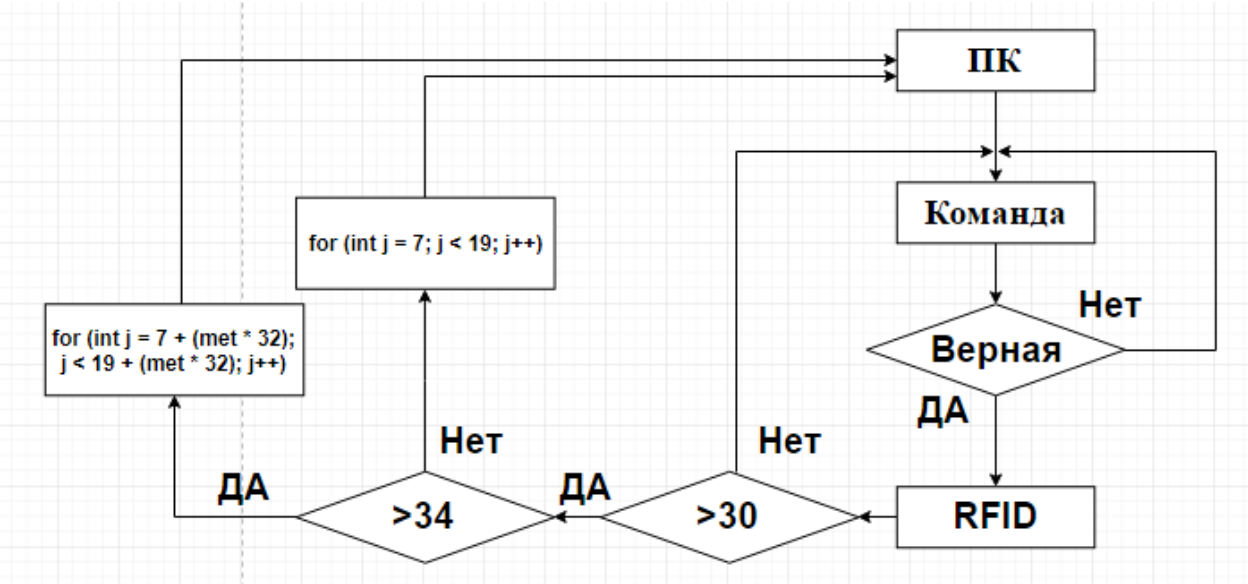

Рисунок 4 – Метод обработки информации.

Достигнуты следующие результаты по поставленным задачам:

1. Разработан алгоритм подключения к RFID-считывателю;
2. Разработан алгоритм формирования запроса на считывание;
3. Разработан алгоритм обработки полученного результата.

В ходе разработки на основание полученных результатов можно сказать, что разработанные алгоритмы могут применяться для идентификации радиочастотных меток на основе RFID-идентификации для построения систем автоматической инвентаризации.

Список использованной литературы

1. Астафьев А.В. Разработка алгоритма прогнозирования и предотвращения нештатных ситуаций в системах контроля движения промышленной продукции на основе анализа данных мультикодовой маркировки. Известия Юго-Западного государственного университета. 2019;23(4):116-128. https://doi.org/10.21869/2223-1560-2019-23-4-116-128
2. Астафьев А.В. Разработка алгоритма позиционирования мобильного устройства на основе сенсорных сетей из BLE-маяков для построения систем автономной навигации / А.В. Астафьев, А.А Демидов, М.В. Макаров, Д.Г. Привезенцев // Всероссийская конференция ММРО-2019. Россия, г. Москва, 26-29 ноября 2019 г. с.334-335.
3. Орлов А.А. Разработка алгоритма определения перемещений изделий между стеллажами на основе данных с их меток / А.А. Орлов, А.В. Астафьев, Д.Г. Привезенцев // Телекоммуникации, Наука и технологии, ISSN: 1684-2588. №1. 2019. С. 7-15.

Данилин С.Н., А.Д. Зуев
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: dsn-55@mail.ru

Уточнение терминов «информация» и «сигнал»

Галушкин А.И. в монографии [1] обратил особое внимание на важность унификации и стандартизации терминов и определений в области теории и практики искусственных нейронных сетей (ИНС) для однозначного их понимания и адекватного применения всеми исследователями, разработчиками, производителями и пользователями. В обзорно – аналитических работах [2-3] и ряде других проблема несовпадения терминологии показана особенно явно.

Авторы доклада выполнили данную работу применительно к понятиям «информация» и «сигнал», как объектам преобразования в искусственных нейронных сетях на базе мемристоров. Сигналом называется «процесс изменения во времени физического состояния какого-либо объекта, служащий для отображения, регистрации и передачи сообщений» [4]. Более кратко: сигнал – носитель информации.

В научно-практической области понятие «информация» является дискуссионным.

Для согласованного в научном сообществе решения актуальных научно-технических проблем, в частности по теории и практике ускоренного создания в России цифровой индустрии [5], необходимо продолжение работ по формированию адекватного объективной реальности представления об информации.

Авторами предложен общий подход [6-8] формирования представления об информации и ее свойствах, который включает в себя следующие положения и допущения:

а) информацию необходимо исследовать классическими философскими методологиями и характеризовать в соответствии с философскими категориями;

б) информацию необходимо изучать и исследовать методологиями, методами, алгоритмами кибернетики и теории информации;

г) регулярный мониторинг результатов новейших научно-технических и фундаментальных исследований;

в) согласованность с действующим российским законом от 27.07.2006. N 149-ФЗ «Об информации, информационных технологиях и защите информации» [9].

На основе общего подхода и новой редакции Закона, сформулировано и обосновано уточненное определение термина «информация»:

«Информация - это мера (сведения, сообщения, данные, параметры, характеристики) качественных и (или) количественных свойств любых объектов или процессов, явлений или событий не зависимо от формы ее представления или существования».

На основе теории системного анализа показаны основные функции, которые выполняет информация в природе и обществе, науке и технике.

Сформулированное представление об информации, позволило авторам разработать общий подход, методы и алгоритмы проектирования, производства, эксплуатации отказоустойчивых и надежных ИНС на базе мемристоров [10].

Работа выполнена при поддержке гранта РФФИ №19-07-01215

Литература

1. Галушкин, А. И. Нейронные сети: основы теории / А. И. Галушкин. – М.: Горячая линия-Телеком, 2013 – 496 с.

2. Torres-Huitzil C., Girau B. Fault and error tolerance in neural networks: A review. // IEEE Access. 2017. V. 5. P. 17322-17341

3. Yeung D. S., Cloete I., Shi D., Ng W. W. Y. Sensitivity Analysis for Neural Networks. Heidelberg: Springer, 2010. P. 89

4. Баскаков С.И. Радиотехнические цепи и сигналы. М.: Высшая школа, 2000. – 462 с.

5. Данилин С.Н., Щаников С.А., Сакулин А.Е. Перспективы применения нейрокомпьютеров для создания цифровых двойников // Нейрокомпьютеры и их применение XVI Всероссийская научная конференция: тезисы докладов. Москва, 2018. С. 143-144.

6. Данилин С.Н. Современное представление об информации // Информационные системы и технологии. 2012. № 4. С. 138-146.

7. Данилин С.Н. О современном понятии информации // Информационные технологии. 2003. №11. С. 52-57.

8. Данилин С.Н., Щаников С.А. Современное представление об информации // XV Всероссийская научная конференция «Нейрокомпьютеры и их применение». Тезисы докладов. – М:ФГБОУ ВО МГППУ, 2017. С. 67-68

9. Федеральный закон от 27.07.2006 N 149-ФЗ (ред. от 02.12.2019) "Об информации, информационных технологиях и о защите информации" (с изм. и доп., вступ. в силу с 13.12.2019). Официальный Интернет-портал правовой информации http://www.pravo.gov.ru.

10. Данилин С.Н., Щаников С.А., Зуев А.Д., Борданов И.А., Сакулин А.Е. Проектирование искусственных нейронных сетей на основе мемристоров с заданной отказоустойчивостью. Радиотехнические и телекоммуникационные системы. 2019. № 2 (34). С. 41-50.

Данилин С.Н., Щаников С.А., Борданов И.А., А.Д. Зуев
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: dsn-55@mail.ru

Классификация основных факторов, снижающих отказоустойчивость и надежность искусственных нейронных сетей на базе мемристоров

Достигнутое на этапе компьютерного моделирования номинальное качество работы искусственных нейронных сетей (ИНС) (в том числе, отказоустойчивость и надежность) значительно снижается при их технической реализации. Причина этого явления заключается в неизбежном влиянии внутренних и внешних физических и информационных дестабилизирующих работу ИНС факторов, а также производственных и эксплуатационных погрешностей значений параметров компонентов структуры и элементов платформы их реализации [1-3].

Все возможные основные факторы, вызывающие изменения отказоустойчивости (ОУ) и надежности (Н) функционирования ИНС на базе мемристоров (ИНСМ) можно условно разделить на внутренние и внешние. Например, внутренние факторы: - эксплуатационные погрешности элементов; - шумы элементов и структур; - ошибки алгоритмического, математического и программного обеспечения. Внешние факторы: - производственные погрешности элементов и структур; - шумы и ошибки во входной информации; - помехи и погрешности источников электропитания ИНСМ. Для обеспечения заданной ОУ и Н функционирования (в зависимости от предъявляемых требований к уровню надежности) необходимо определить и учесть влияние всех существенных внутренних и внешних дестабилизирующих факторов для конкретного варианта реализации ИНСМ. В результате этой работы могут быть определены предельно допускаемые уровни указанных дестабилизирующих факторов (Рис. 1).

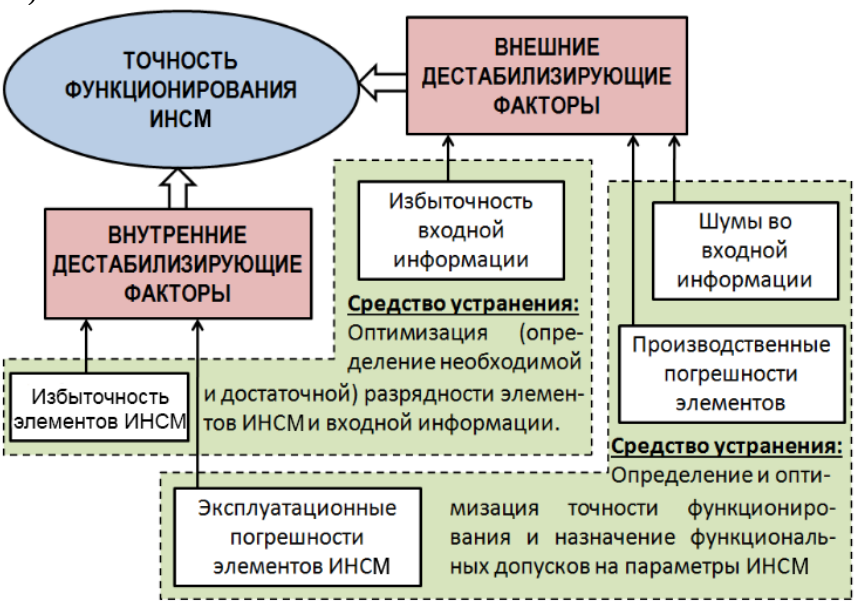

Рис. 1. Факторы, дестабилизирующие работу ИНСМ

На основе обзора научно-технических источников, методологии системного анализа, принципов синтеза и функционирования ИНСМ, результатов имитационного моделирования, авторами предложен следующий вариант классификации основных физических и информационных факторов, снижающих отказоустойчивость ИНСМ [1]:

– внутренние и внешние;

– аппаратные и программные;
– цифровые и аналоговые;
– тип платформы технической реализации;
– структурно–функциональный уровень;
– тип выполняемой функции;
– этап функционирования;
– режим функционирования;
– погрешности и ошибки информационного, математического, методического, алгоритмического и программного обеспечения;
– типы ИНСМ;
– варианты схемотехнических, конструктивных, технологических реализаций;
– погрешности исполнения ИНСМ, управления, эксплуатации;
– чувствительность к внутренним и внешним воздействиям;
– электрические режимы мемристоров (М) и комплектующих элементов (КЭ), характерные для типа элемента, и выполняемых функций.

Общесистемные факторы:
– общие характеристики факторов;
– объективные и субъективные;
– одномерные и многомерные;
– аддитивные и мультипликативные;
– статические и динамические;
– периодические и непериодические;
– линейные и нелинейные;
– зависимые и независимые.

Предложенная классификация позволяет систематизировать дестабилизирующие факторы и более эффективно применять известные и разрабатывать новые подходы, методы, алгоритмы их устранения и (или) компенсации результатов их деструктивного влияния на отказоустойчивости и надежности ИНСМ.

Современные методы и средства активного и пассивного устранения влияния дестабилизирующих функционирование ИНСМ факторов наиболее подробно рассмотрены в работах [2-4].

Работа выполнена при поддержке гранта РФФИ №19-07-01215.

Литература

1. A Survey of Neuromorphic Computing and Neural Networks in Hardware [Электронный ресурс] / C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, J. S. Plank // Proceedings of CoRR journal. URL: https://arxiv.org/pdf/1705.06963.pdf (Дата обращения: 24.08.2019).

2. Torres-Huitzil C., Girau B. Fault and error tolerance in neural networks: A review. // IEEE Access. 2017. V. 5. P. 17322-17341

3. Yeung D. S., Cloete I., Shi D., Ng W. W. Y. Sensitivity Analysis for Neural Networks. Heidelberg: Springer, 2010. P. 89

4. Данилин С.Н. Зуев А.Д. Особенности обеспечения отказоустойчивости нейронных сетей на базе мемристоров на схемотехническом структурно-функциональном уровне. Радиотехнические и телекоммуникационные системы. 2019. № 4 (36). С. 32-43.

Данилин С.Н., Щаников С.А., Борданов И.А., А.Д. Зуев
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: dsn-55@mail.ru

Основные направления повышения отказоустойчивости и надежности искусственных нейронных сетей на базе мемристоров

Актуальность обеспечения необходимой точности функционирования искусственных нейронных сетей (ИНС) при обеспечении высокой отказоустойчивости и надежности показана в ряде работ Галушкина А.И., в частности в монографии [1]. Автор назвал ряд причин, по которым решение этой проблемы является сложным.

В обзорно аналитических работах [2-4], посвященных проблеме обеспечения высокой отказоустойчивости (ОУ) и надежности (Н) ИНС рассмотрены, обобщены, поставлены и решены многие задачи в этой области.

Показано, что работы в области ОУ и Н ИНС проводятся длительное время учеными, теоретиками и практиками. Проблема является сложной, многогранной, с возрастающей размерностью при масштабировании вычислительных структур, и поэтому решается медленно и фрагментарно, требуя применения дополнительных информационных физических, финансовых ресурсов при решении различных типов вычислительных задач.

Главной характеристикой ИНС, как преобразователей информации, является точность функционирования. Другой важной характеристикой ИНСМ является их надежность. Среди характеристик надежности особое место занимает ОУ, так как на основании численных значений ОУ можно построить схему надежности технического устройства и рассчитать основные показатели надежности по стандарту [5].

Такая простая, но эффективная технология, стала возможной после разработки авторами доклада количественной меры ОУ [6].

В работе [3], на большом числе примеров показано, что работы в области повышения ОУ идут в 4-х направлениях:

а) внесение физической или информационной избыточности;

б) внесение физической или информационной автокомпенсации паразитных факторов или результатов их воздействия;

в) применение алгоритмов отказоустойчивого обучения;

г) оптимизация параметров качества по имеющимся ограничениям на проектирование.

Наибольшую результативность удается достигнуть, применяя ОУ обучение, методы которого условно разделяют на 10 подтипов, применяемыми для разных структур ИНС и решаемых ими задач.

В работе [4], глубоко исследованы методы, основанные на снижении чувствительности параметров качества функционирования ИНС к вариациям весов, стабильности функций активации и уровню мультипликативных шумов. Применение этих методов для обеспечения высокой ОУ и Н в ИНС на базе мемристоров (ИНСМ) требует дополнительной проверки в связи с результатами, ранее проведенными авторами доклада работ [7-8]. В данных работах экспериментально установлена непредсказуемая нелинейность функций чувствительности ИНС с произвольной базой реализации в зависимости от значений вариаций весов нейронов.

Работа выполнена при поддержке гранта РФФИ №19-07-01215.

Литература

1. Галушкин А. И. Нейронные сети: основы теории / А. И. Галушкин. Москва: Горячая линия-Телеком, 2013. 496 с

2. A Survey of Neuromorphic Computing and Neural Networks in Hardware [Электронный ресурс] / C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, J. S.

Plank // Proceedings of CoRR journal. URL: https://arxiv.org/pdf/1705.06963.pdf (Дата обращения: 24.08.2019).

3. Torres-Huitzil C., Girau B. Fault and error tolerance in neural networks: A review. // IEEE Access. 2017. V. 5. P. 17322-17341.

4. Yeung D. S., Cloete I., Shi D., Ng W. W. Y. Sensitivity Analysis for Neural Networks. Heidelberg: Springer, 2010. P. 89.

5. ГОСТ Р 27.301–2011. Надежность в технике. Управление надежностью. Техника анализа безотказности. Основные положения. Москва: Стандартинформ, 2013. 19 с.

6. Данилин С. Н., Пантелеев С.В. Алгоритм контроля отказоустойчивости нейронных сетей. //Информационные технологии. №1, 2013 С. 67 – 70. ISSN 1684-6400 .

7. Щаников, С.А. Алгоритм определения коэффициентов влияния погрешностей элементов нейронов на показатели качества работы устройств с нейросетевой архитектурой / С.А. Щаников, С.Н. Данилин, М.В. Макаров // Методы и устройства передачи и обработки информации. – 2011. – №13. – С. 114–118.

8. Щаников, С.А. Исследование коэффициентов влияния погрешностей элементов нейронов на показатели точности (качества) работы устройств с нейросетевой архитектурой [Электронный ресурс] / С.А. Щаников, С.Н. Данилин, М.В. Макаров // Алгоритмы, методы и системы обработки данных. – 2011. – №2(17). – Режим доступа: http://amisod.ru/images/mediacontent/2011/2/ amisod-2011-2-¬17-danilin-makarov-schyanikov.pdf.

Данилин С.Н., Щаников С.А., Борданов И.А., А.Д. Зуев
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: dsn-55@mail.ru

Программно-техническая система для синтеза и анализа искусственных нейронных систем на базе мемристоров

В настоящее время самая известная инструментальная система по исследованию технических средств на базе мемристоров – это ArC ONE [1] производства ArC Instruments. Она содержит полную возможность измерения характеристик мемристоров и обеспечивает формирование входных импульсов с точностью до 10 нс. Амплитуды импульсов до ± 12В с нарастанием / спадом времени вплоть до 30 нс и шириной импульса до 90 нс. Это позволяет охватить широкий спектр вариантов реализаций мемристоров и мемристорных матриц. К основным ограничениям данной системы можно отнести:

а) получение данных лишь о физических параметрах мемристоров. Это является недостаточным для разработки методов определения технических характеристик ИНСМ как физическо-информационных объектов, что показано в ряде работ авторов, в частности [2];

б) проведение исследований с реальными устройствами. Для разработки методов обеспечения надежности функционирования ИНСМ целесообразнее проводить предварительный инженерный расчет допусков, отказоустойчивости, надежности на основе модели [2-5] и только затем проверять эффективность применения данных методов для аппаратного варианта реализации ИНСМ с помощью данной системы;

в) высокую стоимость системы и проведения экспериментов за счет невозможности предварительного контроля точности производимых для исследования образцов НСМ на этапе проектирования.

В рамках задачи импортозамещения, авторами разрабатывается структура и компоненты программно-технической системы для имитационного моделирования, исследования, настройки, управления, контроля как моделей, так и физических реализаций мемристоров, мемристорных матриц, нейронов и ИНСМ для получения априорных и апостериорных данных о значениях их технических параметров, необходимых для выполнения проектирования ИНСМ с требуемой отказоустойчивостью и надежностью. По ряду функциональных возможностей и основных технических показателей она превосходит возможности описанной системы. Разработка и реализация российская, создана в научном коллективе МИ ВлГУ.

Работа выполнена при поддержке гранта РФФИ №19-07-01215.

Литература

1. Serb A., Bill J., Khiat A., Berdan R., Legenstein R., Prodromakis T. Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses // Nature Communications 2016. Vol. 7. doi: 10.1038/ncomms12611.

2. Galushkin A.I., Danilin S.N., Shchanikov S.A. The research of memristor-based neural network components operation accuracy in control and communication systems // 2015 International Siberian Conference on Control and Communications, SIBCON 2015 – Proceedings. 2015. PP. 1-6. doi:10.1109/SIBCON.2015.7147034.

3. Danilin S.N, Shchanikov S.A., Panteleev S.V. Determining Operation Tolerances of Memristor-Based Artificial Neural Networks // Proceedins of the 2016 International Conference on Engineering and Telecommunication (EnT-2016). 2016. PP. 33-37.

4. Danilin S.N., Shchanikov S.A. The research of operation accuracy of a memristor-based artificial neural network with an input signal containing noise and pulse interference // Proceedings - X International IEEE Scientific and Technical Conference Dynamics of Systems, Mechanisms and Machines (Dynamics). 2016.

5. Danilin S.N., Shchanikov S.A. Neural Network Control Over Operation Accuracy of Memristor-based Hardware // Proceedings of 2015 International Conference on Mechanical Engineering, Automation and Control Systems, MEACS 2015. 2015. PP. 1-5. doi:10.1109/MEACS.2015.7414916

Канунова Е.Е.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: kanunovaee@list.ru

Автоматизация документооборота в средней общеобразовательной школе

Основные задачи делопроизводства в образовательном учреждении — это сокращение информационных потоков до оптимального минимума и обеспечение упрощения и удешевления процессов сбора, обработки и передачи информации с помощью новейших технологий автоматизации этих процессов.

В образовательном учреждении принимаются, формируются, согласовываются и исполняются следующие виды документов:

- входящие документы – письма, распоряжения вышестоящих органов, заявления;
- исходящие документы – отчеты, характеристики, справки письма;
- внутренние документы – журналы, расписания, приказы и распоряжения, локальные акты.

Движение документов осуществляется как по внешнему, так и по внутреннему контуру (рисунок 1). В работе школьного документооборота принимают участие следующие типы сотрудников: секретари, руководители и исполнители.

Относительно конкретного исполнителя все документы, с которыми он работает, делятся на несколько категорий:

- входящие, с которыми исполнитель не успел ознакомиться;
- в работе, которые ждут его действий;
- на контроле, по которым он ожидает действий от других исполнителей.

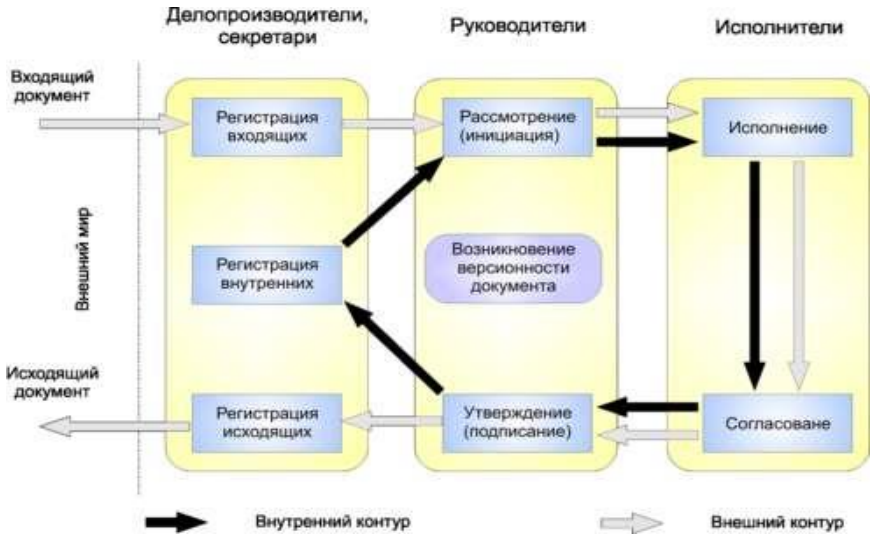

Рисунок 1 – Структура документооборота школы

Если документ создается внутри школы, то он считается внутренним, и для его оборота формируется внутренний контур, включающий следующую последовательность действий: регистрация внутреннего документа секретарем, рассмотрение его руководителем, исполнение сотрудником школы, выполняющем роль исполнителя, также согласование исполнителем и утверждение руководителем.

Документ, поступающий в школу из других организаций, принимает статус внешнего, и его движение организуется по внешнему контуру. Внешний контур включает дополнительный этап, который выполняет секретарь – регистрация исходящего документа. После чего внешний документ принимает статус исходящего и передается в другие организации.

Основная роль в организации документооборота лежит на секретаре школы. К основным функциям секретаря школы относятся:

- прием и регистрация входящих документов, как внешних, так и внутренних.
- передача документов на рассмотрение руководству и их получение от него с указаниями и резолюциями о порядке исполнения этих документов;
- прием и отправка исходящих документов;
- прием документов от работников школы для передачи на рассмотрение директором, для изготовления документов по установленной форме, для копирования и размножения;
- передача работникам школы документов поступивших от руководства с резолюциями, поступивших после изготовления, копирования;
- контроль за исполнением документов и поручений руководства;
- оформление и хранение документов, а также печатей, штампов и бланков;
- составление и изготовление документов в соответствии с поручениями и указаниями руководства, изготовление и выдача копий;
- обеспечение сохранности документов.

Задачи руководителя, в роли которого выступает директор школы или заместители директора, рассмотрение зарегистрированного секретарем документа, назначение исполнителей документов и подписание документа. Если имеет место факт назначения документа нескольким исполнителям, то руководитель контролирует версии документа.

В докладе рассматриваются особенности разработки и использования системы автоматизации документооборота в общеобразовательном учреждении на примере средней школы №4 города Мурома.

Разработанное приложение автоматизирует все задачи, связанные с документооборотом в школе и реализует следующие ключевые функции:

- регистрация пользователей в системе;
- формирование и хранение карточки документа;
- согласование документа пользователями системы, которые являются участниками согласования;
- назначение исполнителей для каждого документа в системе;
- ведение журнала событий, происходящих с документом, с момента его создания;
- контроль за исполнением указаний, описанных в документе, в указанный срок;
- формирование отчетов, связанных со статистикой поступления, исполнения, согласования документов.

Система реализована в виде одностраничного web-приложения, т.е. приложения, использующего единственный HTML-документ как оболочку для всех web-страниц и организующего взаимодействие с пользователем через динамически подгружаемые HTML, CSS, JavaScript посредством асинхронного обмена данными с сервером. Программная архитектура системы автоматизации документооборота школы построена в соответствии с паттерном MVC (Model (модель)-View (представление)-Controller (контроллер)). В качестве средств реализации использовалась среда разработки Microsoft Visual Studio 2017, язык C#, технология ASP.NET MVC 5, .NET Framework 4.5, платформа разработки Entity Framework 6.3, СУБД Microsoft SQL Server Express 2016.

Использование разработанной информационной системы автоматизации документооборота позволит повысить оперативность работы сотрудников школы при решении задач работы с документами, как внутренними, так и внешними.

Литература

1. Федеральный закон «Об образовании в Российской Федерации» // URL: http://www.consultant.ru/document/cons_doc_LAW_140174 (дата обращения 09.01.2020

Канунова Е.Е.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: kanunovaee@list.ru,

Сентиментальный анализ текста с использованием нейросетевых моделей

Актуальной задачей является задача определения тональности текстов. Например, на основе результатов анализа отзывов пользователей сети Интернет компания, предоставляющая те или иные услуги, может выработать политику деятельности, приносящую прибыль. Существуют и более глобальные задачи, например, исследование политических настроений в преддверии выборов или оценка существующей власти.

Любой текст несет в себе ту или иную эмоцию - радость, гнев, страх и т.п. Эмоцию можно классифицировать как позитивную и негативную без уточнений. Разработка подхода к определению эмоции в тексте позволит ускорить процесс его сентиментального анализа. Это в свою очередь даст возможность оперативно отреагировать на те или иные высказывания людей.

В докладе представлены результаты исследований методов глубинного обучения и их применение к задачам обработки текстов. Приводится обзор методов обучения, основанных на применении многослойных нейронных сетей [1]. Приводятся результаты определения тональности текста на основе модели Word2Vec [2].

При реализации проекта использовался язык программирования Microsoft Visual C# и библиотека классов Microsoft.ML. Описаны принципы создания и обучения модели нейронной сети с помощью Framework Microsoft.ML.NET. Представлено приложение для работы с созданной моделью для анализа тональности текста. В качестве данных для обработки были использованы отзывы покупателей, взятые с ресурса [3] и показаны результаты его работы.

Литература

1. Многослойная нейронная сеть // URL: https://neuralnet.info (дата обращения: 08.01.2020)
2. Модель Word2Vec // URL: https://pathmind.com/wiki/word2vec (дата обращения: (08.01.2020)
3 Интернет-магазин Эльдорадо // URL: https://eldorado.ru (дата обращения: 08.01.2020)

Косаурова А.Д., Андрианова В.И.
Владимирский государственный университет им. А. Г. и Н. Г. Столетовых
600000, г. Владимир, ул. Горького, 87
E-mail: anakos1996@gmail.com, mazanova_v@mail.ru

Оценка внедрения готового решения на платформе «1С: Предприятие 8»

В условиях цифровой экономики вопрос подбора соответствующего программного обеспечения является актуальным. Конечно, в первую очередь необходимо учитывать функционал программы, направленный на взаимодействие с клиентом. Поэтому компании стремятся подобрать хотя бы два наиболее подходящих решения, а уже из них выбрать оптимальный вариант по критерию стоимости внедрения системы.

Для оценки стоимости внедрения можно использовать методику ТЭО (Технико-экономическое обоснование), которая применяется в проектах, специализирующихся на разработке программного обеспечения [1]. Данная методика используется, чтобы:

- определить целесообразность проекта по разработке и внедрению информационной системы;
- рассчитать и проанализировать затраты;
- оценить сроки окупаемости системы.

Методика ТЭО адаптирована для оценки внедрения готового решения на платформе «1С: Предприятие 8». Для проведения технико-экономического обоснования необходимо учитывать затраты на:

- описание задачи ($Т_о$);
- затраты на закупку ($Т_з$);
- программирование (доработку) ($Т_п$);
- отладку ($Т_{отл}$);
- затраты на обучение персонала ($Т_{об}$);
- перенос данных из текущей системы в новую ($Т_{пд}$).

Итоговые затраты на внедрение готового программного обеспечения ($Т_{впо}$) рассчитываются по формуле (1).

$$Т_{впо} = Т_о + Т_з + Т_п + Т_{отл} + Т_{об} + Т_{пп} \quad (1)$$

Для большинства критериев в оценке трудозатрат используется формула 2, где D – стоимость трудозатрат, m – стоимость часа работ, t – необходимое время реализации в часах, c – техническая сложность задачи (целое положительное число).

$$D = mtc \quad (2)$$

Каждый из критериев затрат рассчитывается по соответствующей формуле.

1) Описание задачи рассчитывается по формуле 2, техническая сложность задачи при этом равна единице.

2) В затраты на закупку входит стоимость самого программного обеспечения ($Т_{по}$), стоимость лицензий ($Т_л$), стоимость закупки аппаратного обеспечения ($Т_{ао}$). Таким образом затраты на закупку рассчитываются по формуле 3.

$$Т_з = Т_{по} + Т_л + Т_{ао} \quad (3)$$

Затраты на программирование (доработку) учитывают в себе стоимость, рассчитанную по формуле 2, но добавляют в себя затраты на среду разработки ($Т_{ср}$). Следовательно, затраты на доработку можно рассчитать по формуле 4.

$$Т_п = D + Т_{ср} \quad (4)$$

3) Затраты на отладку включают в себя половину затрат на разработку без учета стоимости среды разработки. Их можно рассчитать по формуле 5.

$$Т_{отл} = 0{,}5D \quad (5)$$

4) При затратах на обучение персонала необходимо учесть время, которое потребуется на обучение (t) и среднюю зарплату в час сотрудника, который будет проводить это обучение ($m_{ср}$). Полная стоимость обучения может быть рассчитана по формуле 6.

$$Т_{об} = tm_{ср} \quad (6)$$

5) При переносе данных из действующей информационной системы в новую в первую очередь учитывается сложность данных (с), в которую входит объем данных и сила их взаимосвязанности между собой. Технический перенос данных составляет лишь треть стоимости. Необходимо проверить корректность переноса этих данных. Стоимость переноса данных можно рассчитать по формуле 7, адаптировав формулу 2.

$$T_{пд} = 1{,}5D \quad (7)$$

На рис. 1 приведена структура оценки затрат внедрения готового решения по методике технико-экономического обоснования.

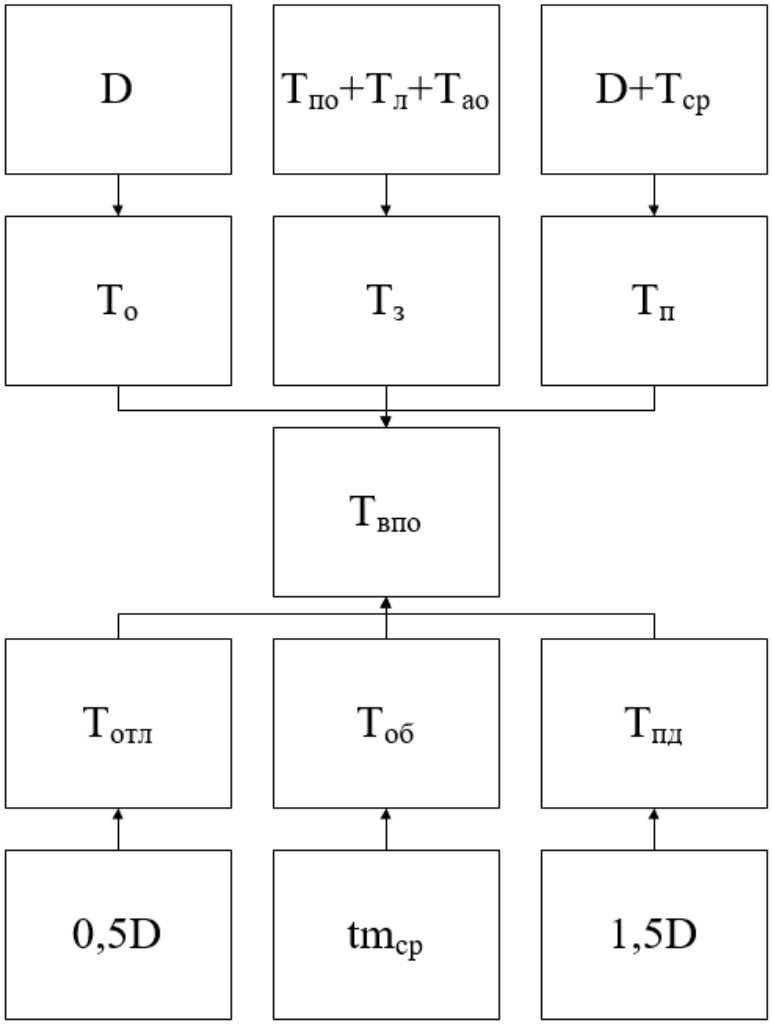

Рис. 1 – Структура оценки затрат

Данная методика может быть использована при оценке внедрения готового программного продукта для любого предприятия, работающего в сфере информационных технологий.

Литература

1. Коцюба И.Ю., Чунаев А.В., Шиков А.Н. Методы оценки и измерения характеристик информационных систем. Учебное пособие. – СПб: Университет ИТМО, 2015. – 264 с.

Родионова А.В.
Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых
г. Владимир, ул. Горького 87
nasty.rod@yandex.ru

Методы распознавания жестов

Очень часто жестикуляция играет не последнюю роль в общении людей, во взаимодействии одного собеседника с другим. Жесты несут в себе определенную часть информации, которую можно проанализировать и использовать. В информационном пространстве жесты стали альтернативой инструментам взаимодействия пользователя и технического устройства. Именно поэтому в настоящий момент исследование и разработка человеко-машинного взаимодействия на основе распознавания образов и визуальном представлении информации является одним из главных направлений в развитие современного программного обеспечения. Использование жестов выглядит особенно многообещающим в задачах построения различных интерфейсов управления программным и аппаратным обеспечением.

Способы распознавания жестов

Вся задача распознавания сводится к двум проблемам: определение объекта (руки) и определение жеста. Существуют различные методы обнаружения руки в пространстве, а, соответственно, и распознавания жестов. Эти методы делятся на две группы: методы на основе внешнего вида (Vision-based-approach) и поиска эталонного образца объекта, методы на основе 3D-модели объекта (3D hand-model-based-approach). Далее будут рассматриваться примеры различных методов из каждой группы

Методы на основе внешнего вида руки

Отличительной чертой методов, базирующихся на анализе внешних признаков жеста, является оценка внешнего вида (формы, расположения и т.д.) целевого объекта. Для распознавания не хранится никакой информации о физических свойствах рассматриваемого объекта.

Одним из примеров таких методов может служить алгоритм распознавания с помощью модели разметки. Данную систему распознавания для мобильных устройств совсем недавно разработали инженеры Google[2].

Новая технология Google включает в себя три модели искусственного интеллекта, работающих во взаимосвязи: детектор ладони, модель для разметки ладони и классифицирующий модуль. Детектор ладони анализирует кадр и возвращает ограничивающий руку прямоугольник, модель для разметки ладони оценивает область изображения и возвращает набор трёхмерных точек, а модуль классифицирует ранее полученную конфигурацию из точек и сопоставляет их с тем или иным жестом.

Еще один метод распознавания – метод Виолы-Джонса, предложенный в 2001 году. И хотя основной задачей алгоритма было распознавание лиц, он стал широко применяться для распознавания объектов на изображениях.

Метод использует технологию скользящего окна [3]. Если кратко, то есть рамка, размером, меньшим, чем исходное изображение, которая двигается с некоторым шагом по изображению, и с помощью каскада слабых классификаторов определяет, есть ли в рассматриваемом окне лицо. Данный метод зарекомендовал себя как достаточно эффективный для задач компьютерного зрения и распознавания объектов.

Также популярными можно назвать алгоритмы, использующие в качестве отличительного признака цвет.

Так, цвет кожи служит признаком для эффективной локализации и отслеживания частей человеческого тела. Именно на данном признаке основан алгоритм распознавания, представленный в работе [4].

Для определения жестов руки могут применяться цветные перчатки [5]. Подобный способ позволяет с помощью одной лишь видеокамеры в реальном времени распознавать конфигурацию руки и отслеживать движения ладони в трехмерном пространстве.

Метод распознавания с помощью маркеров очень похож на метод, использующий цветные перчатки. Маркерная система использует специальное оборудование. Для захвата движения на руку человека прикрепляются специальные датчики, которые считывают и передают данные об изменении положения руки. Далее на основе этих данных строится трехмерная модель, точно воспроизводящая жесты человека, и, на основе уже этой модели создается анимация жестов. Исследователи из Массачусетского Технологического Института долгое время занимаются исследованиями, связанными с распознаванием жестом, поэтому вполне можно привести в пример их разработки. Для построения модели руки на основании изображения ими было предложено использовать в качестве маркеров разноцветные перчатки, что значительно упрощало задачу и позволяло производить распознавание практически в режиме реального времени.

Здесь метод распознавания на основе маркеров похож на метод распознавания с помощью цветных перчаток. Но на самом деле использование цветных перчаток является частным случаем метода распознавания по маркерам. Маркеров могут выступать в различной конфигурации и использоваться как по всему телу, так и на его различных участках. Обычно маркеры представляют собой небольшие шарики из светоотражающего материала. Маркеры крепятся на местах сгиба конечностей: пальцы, кисти, локти и т.д. Компьютер считывает их расположение на «скелете», а далее процесс дорисовки становится намного проще. Данный способ широко используется в кинематографе[12].

Системы на основе захвата движения при помощи маркеров ориентированы на точность за счет простоты использования и установки.

Популярным также является метод распознавания с помощью моментов изображения. Моменты изображения при некоторых ограничениях могут использоваться для распознавания простых жестов рук и создания на их основе приложений HCI. Например, в работе [6] рассматривается программа, позволяющая управлять игрушечным роботом при помощи руки, где ориентация кисти определяет направления движения робота.

В распознавании важно отделить анализируемый объект от лишних предметов. Так, на практике удовлетворять условиям однородного фона удается не всегда. Для таких случаев тоже существует свой метод. Он основан на анализе центра массы разностей изображений (motion energy image - MEI) руки в кадрах видеоряда [7].

Методы на основе 3D моделей

Технология анализа трехмерной модели руки используется в компьютерном зрении для распознавания детальной трехмерной конфигурации руки при наличии на входе одного или нескольких изображений жеста. Под детальной конфигурацией следует понимать позицию, ориентацию ладони и ключевых точек пальцев руки в трехмерном пространстве.

В большинстве случаев приходится распознавать ограниченное количество жестов. Исходя из этого, многие исследователи создают трехмерную модель руки не во время распознавания, а во время обучения системы. В работе [8] создается система, которая умеет распознавать 24 жеста. Для каждого из 24 жестов храниться 15 изображений руки, проектируемых из разных углов наблюдений. Вместе с изображениями хранятся параметры конфигурации трехмерной модели руки.

Оптимизация требуется везде. Задачи оптимизации решаются и для методов распознавания. Метод роя частиц [9] как раз-таки может выступать в качестве одного из способов оптимизации алгоритмов распознавания жестов.

В [11] представлен метод, синхронизирующий виртуальную руку, которая представляет собой набор примитивов, с реальной рукой. Виртуальная рука состоит из 27 параметров и задача сводится к подбору значений этих параметров так, что бы они наиболее точно отражали жест реальной руки. В основе как раз лежит метод роя частиц (МРЧ).

Сравнение и выводы

Таким образом, были описаны одни из самых распространенных и популярных на сегодняшний день методов, используемых для распознавания руки человека. Ниже приведена сравнительная таблица данных алгоритмов по некоторым характеристикам.

Таблица 1. Сравнение алгоритмов распознавания жестов.

	Распознавание с помощью модели разметки	Метод Виолы-Джонса	Распознавание с использованием цветных перчаток	Моменты изображения	MEI	Метод роя частиц
Точность распознавания простых жестов	86-94%	90%	80-90%	92-94%	80-90%	90%
Дистанция работы метода	до 20м	1-2м	1-3м	достаточно большое, до 1000м	1-3м	1-4м
Производительность (кол-во кадров в сек.)	~21	~15	6-10	20-30	~30	~8
Требование маркеров или специальных перчаток	нет	нет	да	нет	нет	нет

Из таблицы 1 видно, что все алгоритмы имеют достаточно высокую точность распознавания, но далеко не стопроцентную. Это значит, что задача распознавания жестов все еще остается весьма актуальной и современные алгоритмы распознавания требуют совершенствования и оптимизации. Для этого нужно добиться решения ряда проблем, которые мешают распознаванию улучшению качества распознавания, например, такие проблемы, как сложность обучения или разная освещенность окружающей среды при распознавании.

Литература

[1] Нагапетян В.Э., Методы распознавания жестов руки на основе анализа дальностных изображений // РУДН, 2013 г.

[2] Тверье С., Инженеры Google создали систему для распознавания жестов для мобильных устройств, - статья. 21.08.2019, - доступ: http://citforum.ru/news/40384/

[3] P. Viola и J. Michael , «Computer Vision and Pattern Recognition,» в Rapid Object Detection using a Boosted Cascade of Simple, 2001

[4] WU Y., HUANG T. S. Non-stationary color tracking for vision-based human computer interaction // IEEE Trans. Neural Networks. 2002. V. 13, N 4. P. 948– 960.

[5] Wang R.Y., Popovi´c J. Real-time hand-tracking with a color glove // ACM Trans. Graph., Volume 28, Number 3, ACM: 2009. — p. 63:1 – 63:8

[6] Freeman W.T., Anderson D.B., Beardsley P., Dodge C.N., Roth M., Weissman C.D., Yerazunis W.S., Kage H., Kyuma K., Miyake Y., Tanaka K.I. Computer vision for interactive computer graphics // Computer Graphics and Applications, IEEE , vol.18, no.3, 1998. — p. 42 – 53

[7] Bobick, A., Davis, J. An appearance -based representation of action. International Conference on Pattern Recognition, pp. 307-312, 1996.

[8] Tomasi C., Petrov S., Sastry A. 3D tracking = classification + interpolation. in Proc. International Conference on Computer Vision, 2003. — p. 1441–1448.

[9] Kennedy, J.; Eberhart, R. (1995). "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks IV, pp.1942-1948

[10] Алгоритм роя частиц. Описание и реализации на языках Python и C#, доступ: https://jenyay.net/Programming/ParticleSwarm

[11] Project Report, December 18, 2003. I. Oikonomidis, N. Kyriazis, and A. Argyros, “Efficient model-based 3D tracking of hand articulations using Kinect”, in BMVC 2011, 2011

[12] Волошин В., Секрет реалистичности кат-сцен Death Stranding, доступ: https://trashbox.ru/link/death-stranding-cut-scenes-secret, - 14.11.2019

Шарапова Е.В.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Обнаружение полных дубликатов текстов

Полный дубликат документа — это экземпляр текстового документа на естественном языке, без учета формы представления (оформления) и формы хранения (формата файла).

Примерами полных дубликатов являются произведения одних и тех же авторов, напечатанные в разных издательствах, тексты, сохраненные в разных текстовых форматах (DOC, DOCX, RTF, TXT), различных кодировках (cp866, windows 1251, KOI8-R, Unicode).

Для обнаружения полных дубликатов наиболее подходит использование сигнатурных методов. Суть сигнатурных методов сводится к представлению документа неким кодом (хэш-функцией, числом, контрольной суммой), позволяющим с высокой степенью вероятности выявить одинаковые (или практически одинаковые) документы. Фактически, сравнение документов сводится к сравнению нескольких чисел — сигнатур документов. Это существенно сокращает потребности памяти и вычислительные затраты. Отличительная особенность сигнатур — это возможность подсчета их в любое время, в том числе заранее, а не в момент проверки.

Перед подсчетом сигнатур предлагается проводить предварительную обработку текстов: приведение к единой кодировке, удаление переносов и разрывов страниц, склейка разорванных при этом слов, удаление номеров страниц, очистку форматирования и представление документов в виде обычных текстов.

Простейшим примером сигнатуры является контрольная сумма документа, например, CRC32 или MD5.

Один из способов подсчета сигнатур [1, 2] основан на подсчете частоты встречаемости слов в документе TF (term frequency). Сигнатура строится по нескольким (5 или 6) наиболее частотным словам. В качестве сигнатуры используется контрольная сумма CRC32 строки, состоящей из выбранных слов, упорядоченных по алфавиту.

Немного более сложная версия сигнатуры предполагает подсчет веса слов не по формуле TF, а по формуле TF*IDF. В данном случае учитывается не только частота слов в документе (TF), но и общая встречаемость слов во всех документах коллекции (IDF). Для подсчета сигнатуры, подсчитывается вес каждого слова, путем вычисления tf(d)×idf(t) и выбираются несколько слов с наибольшим весом. Сигнатура вычисляется как контрольная сумма CRC32 строки, состоящей из выбранных слов, упорядоченных по алфавиту.

Хорошо зарекомендовала себя сигнатура, построенная на основе комбинации частоты слов TF и остаточной обратной частоты документов RIDF. Для подсчета сигнатуры, аналогично предыдущему способу, подсчитывается вес каждого слова по формуле tf(d)×ridf(t), выбираются несколько слов с наибольшим весом, слова упорядочиваются по алфавиту и сцепляются в строку, для которой вычисляется контрольная сумма CRC32.

Сигнатура $TF*IDF_{Opt}$ является модифицированной версией сигнатуры TF*IDF. Модификация заключается в изменении принципа подсчета значения IDF на основе так называемой «оптимальной частоты».

Сигнатура, построенная на основе двух самых длинных предложений, позволяет довольно хорошо находить похожие документы [1]. Для этого в тексте находятся два самых длинных предложения и сцепляются в одну строку в алфавитном порядке. Для строки подсчитывается контрольный код CRC32, который и является сигнатурой. Надо заметить, что при неправильном определении границ предложений (например, при переносах на другую страницу или для больших списков) сигнатура может не определить похожие документы.

Сигнатура двух самых «тяжелых» предложений строится по аналогичному принципу [1]. Из текста выбираются два предложения. Но предложения выбираются на основе суммы весов

(рассчитываемых по формуле tf(d)×idf(t)) входящих в нее слов. Два предложения с наибольшей суммой весов упорядочиваются по алфавиту, сцепляются в одну строку, для которой подсчитывается контрольный код CRC32.

Сигнатура I-Match строится на основе вычисления значения функции, предложенной в [3]. Для всей коллекции документов составляется словарь (список) слов, имеющих среднее значение IDF (слова со слишком большими или маленькими значениями IDF в список не включаются). Для каждого документа формируется множество слов и определяется его пересечение со словарем. При превышении пересечением некоторого порога, для множества слов вычисляется хеш-функция SHA1 (I-Match сигнатура).

Рассмотренные сигнатуры позволяют представить документ одним числовым значением. Понятно, что чем меньше сигнатура, тем менее она подходит для определения нечетких дубликатов текстов. Причина проста — если изменения в тексте сказываются на результирующем значении сигнатуры, то оценить меру похожести текстов уже не удастся. Решением проблемы может являться расширение сигнатуры, то есть включение в нее ограниченного ряда значений неких показателей. В этом случае, похожесть документов может определяться как некая мера (функция) близости двух сигнатур. Чаще всего это совпадение части значений сигнатур похожих документов [4].

Надо заметить, что длина сигнатур должна быть, по возможности, фиксированной и не превышать нескольких килобайт. Дело в том, что при различной длине сигнатур усложняется функция вычисления их близости, а при большой длине – повышается трудоемкость вычислений.

Для вычисления сигнатуры MegaShingles для всего множества шинглов документа производится вычисление 84 различных хэш-функций. Далее по критерию максимума или минимума каждой функции выбираются 84 шингла, которые разбиваются на 6 групп, по каждой из которых строятся 6 супершинглов. Сигнатура состоит из 15 чисел (мегашинглов), представляющих собой все возможные парные сочетания указанных 6 супершинглов.

Неплохие результаты показала сигнатура, построенная на основе словаря опорных слов [5]. Суть ее сводится к следующему. Из всего словаря, построенного по коллекции, выбираются несколько тысяч слов, наиболее качественно описывающих документы. Для каждого документа высчитывается бинарный вектор, длина которого равна количеству опорных слов. Для каждого опорного слова подсчитывается значение TF. В случае, если значение превышает некоторый порог, в соответствующую позицию вектора записывается 1, в противном случае 0. Сигнатура документа представляет собой указанный бинарный вектор. В отличие от ранее описанных сигнатур, мера схожести документов вычисляется как мера совпадения бинарных векторов (при этом совпадение не обязательно должно быть полным). Сигнатура отличается большой длиной. Так, для 4000 опорных слов длина сигнатуры составляет 500 байт. Фактически сигнатура является результатом фиксации векторной модели по длине словаря. Тем не менее, она сохраняет суть сигнатур — представление документа числовым значением фиксированной длины

Так как различные документы могут содержать сноски, комментарии, колонтитулы и т.д. использование одной сигнатуры, например, контрольной суммы, может не дать должного результата. По этой причине предлагается использовать наборы сигнатур, включающие контрольные суммы документов, сигнатуры на основе наиболее частотных слов, сигнатуры на основе самых длинных или самых значимых предложений.

Такое сочетание позволяет нивелировать влияние различных факторов и обеспечивает хорошую полноту поиска полных дубликатов.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №19-07-00692

Литература

1. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007: сб. работ участников конкурса –Переславль-Залесский, 2007. – Т. 1. – С. 166-174.

2. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации // Проблемы управления и моделирования в сложных системах: Труды

XXI Международной конференции (3-6 сентября 2019 г. Самара, Россия). – Самара: ООО “Офорт”, 2019. Том 2. С. 335-339.

3. Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. – P. 605-610.

4. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации с помощью сигнатур содержания // Управление развитием крупномасштабных систем (MLSD’2019) [Электронный ресурс] : материалы Двенадцатой междунар. конфер, 1–3 окт. 2019 г., Москва. – М.: ИПУ РАН, 2019. С. 1009-1011.

5. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference 2002.

Шарапова Е.В.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Структура поискового индекса системы обнаружения нечетких дубликатов текстов

При построении систем обнаружения дубликатов текстов одним из ключевых параметров, влияющим на производительность, является организация поискового индекса. Значение имеют такие параметры, как размеры индекса и скорость доступа к данным.

Исследования показали, что использование в качестве поискового индекса различного рода СУБД (MySQL, Oracle, MSSQL Server и т.д.) для больших коллекций документов не обеспечивает требуемого уровня производительности - время доступа к данным достаточно большое.

Для хранения поисковых индексов могут использоваться текстовые или структурированные бинарные файлы. При хранении индекса в виде текстовых файлов с индексом легко работать, но обработка проводится построчно. Это замедляет скорость работы с индексом. Во втором случае можно добиться высокой скорости операций чтения индекса и перемещения по нему. По этой причине использование структурированных бинарных файлов для хранения поискового индекса кажется наиболее привлекательным вариантом.

Рассмотрим структуру поискового индекса, применяемого в система Автор.NET более подробно. Поисковый индекс можно разделить на несколько групп файлов:

1. Термы.

В группу входит словарь термов (слов, встречающихся в документах), список стоп слов (наиболее часто встречающихся слов из словаря, которые являются не информативными и не рассматриваются при поиске), IDF индекс.

IDF индекс представляет собой значения инверсных частот документов, подсчитанных для каждого терма (слова). Инверсных частот вычисляется по формуле [1]:

$$\mathrm{idf}(t) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

где $|D|$ – число документов в коллекции, $|\{d_i \in D | t \in d_i\}|$ – число документов, в которых встречается слово t.

2. Документы.

Группа включает в себя список всех внесенных в систему документов, с указанием их уникальных идентификаторов, расположением в системе и сети Интернет, языка документа, числа слов в нем, принадлежность к той или иной коллекции и т.д., а также полные тексты документов в текстовом формате в кодировке Unicode.

3. TF*IDF индекс.

Группа включает в себя индексы значений частоты термов TF и TF*IDF, подсчитанные для каждого слова в каждом документе. Значение TF вычисляется по формуле:

$$\mathrm{tf}(t) = \frac{n_t}{\sum_k n_k}$$

где n_t – число вхождений слова t в документ, $\sum_k n_k$ – общее количество слов в документе.

Значение TF*IDF представляет собой произведение значений tf(d)*idf(t).

4. Сигнатуры.

В группу входят сигнатуры, подсчитанные для каждого документа – контрольная сумма CRC32 и MD5, сигнатуры шести самых значимых слов по значениям TF, TF*IDF, TF*RIDF [2, 3], сигнатуры самых длинных и самых значимых предложений, а также «длинная» сигнатура опорных слов [4].

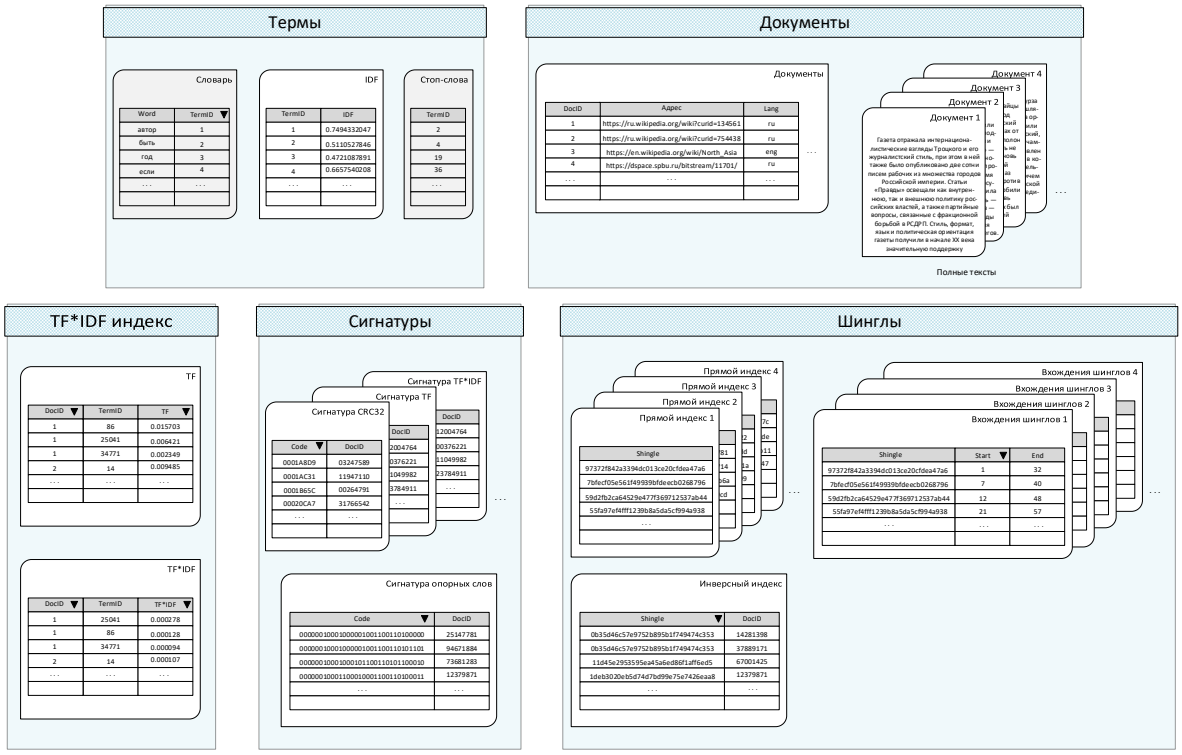

Рис. 1. Структура поискового индекса.

5. Шинглы.

Группа включает в себя наборы прямых индексов шинглов [5] (построенная по порядку следования щинглов в документе) для каждого документа, индексы вхождения шинглов в каждом документе (начальная и конечная позиция в тексте), а также инверсный индекс, сортированный по значениям шинглов.

Предложенная структура позволяет разбить индекс на независимые части и работать с каждой из их по отдельности, в зависимости от выполняемых задач. Кроме того, размеры индексов сигнатур достаточно малы, что позволяет хранить их в оперативной памяти для обеспечения быстрой обработки.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №19-07-00692

Литература

1. Sharapova E.V., Sharapov R.V. Detection of Fuzzy Duplicate Texts in News Feeds // 2019 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Russia, 2019, pp. 1-5.

2. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007: сб. работ участников конкурса –Переславль-Залесский, 2007. – Т. 1. – С. 166-174.

3. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации // Проблемы управления и моделирования в сложных системах: Труды XXI Международной конференции (3-6 сентября 2019 г. Самара, Россия). – Самара: ООО “Офорт”, 2019. Том 2. С. 335-339.

4. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference 2002.

5. Broder A., Glassman S., Manasse M., Zweig G. Syntactic clustering of the web // Computer Networks and ISDN Systems, 1997, vol. 29, n. 8, p. 1157–1166.