

Варламов А.Д., Варламова 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: varlamov_aleks@mail.ru, karetko22@mail.ru

Анализ подходов к оценке схожести фраз в фонетических упражнениях интерактивных онлайн-курсов по изучению английского языка

При изучении английского языка большой проблемой для новичков становится тренировка произношения. Поэтому в состав интерактивных онлайн-курсов по изучению английского языка должны входить фонетические упражнения на произношение английских фраз.

Для создания школьных тренировочных упражнений и тестов на проверку произношения речи на иностранном языке необходимы эталонные фразы, записанные диктором. С данными фразами будут сравниваться записанные в файл высказывания учеников. Таким образом, задача сводится к сравнению двух звуковых файлов. С технической точки зрения реализация данного механизма является сложной задачей. Для ее решения можно использовать следующие подходы:

1. Распознавания произносимой фразы (когда произнесенная фраза переводится в текстовую распознанную строку и затем выполняется сверка текстовых строк).
2. На основе методов сопоставления частотных характеристик сигналов.

Первый подход может быть реализован на основе методов и алгоритмов распознавания речи, которые в настоящий момент являются достаточно развитыми. Но с точки зрения фонетики данный подход является очень грубым: шкала оценки будет двухступенчатая (либо трех или четырехступенчатая в случае длинных фраз), поэтому только очень существенные и критические ошибки произношения будут зафиксированы системой. Подход не будет способствовать выработке учеником отличного безакцентного произношения.

Для реализации второго подхода чаще всего используется метод сопоставления спектрограмм. Для построения спектрограмм в настоящий момент также существуют достаточно эффективные алгоритмы и методы. Но сложности возникают при учете специфики задачи сравнения фраз, которые создают ряд проблем, которые можно охарактеризовать следующим образом:

1. тембр голоса человека, диктовавшего эталонные фразы, может отличаться от тембра голоса ученика;
2. необходима синхронизация по времени сравниваемых фраз, так как спектрограммы имеют временное измерение;
3. необходимо нормирование результата к некоторой оценочной шкале.

Авторы предлагают использовать следующие решения обозначенных проблем:

1. Выполнение частотного сдвига голоса тестируемого человека для уравнивания средних мел частот звуков перед сравнением фраз с целью обеспечения инвариантности алгоритма к тембру голоса.

2. Реализация обрезка звуков таким образом, чтобы в начале и в конце обоих сравниваемых фраз была тихая короткая пауза со строго заданным временным интервалом. Это позволяет исключить возможность получения высокой оценки испытуемым при произнесении им части эталонной фразы вместо полного текста.

3. Использование коэффициента корреляции коэффициентов спектрограммы, который уже сам по себе является нормированной величиной.

Полезность данных решений показали проведенные экспериментальные исследования.

Литература

1. Зотеева Т.С. Использование компьютерных технологий при обучении фонетике английского языка // Педагогика и методика преподавания: цели, технологии, результаты.

сборник научных трудов по материалам I Международной научно-практической конференции. Научно-издательский центр «Открытое знание». 2017. С. 34-40.

2. Хейн М.З., Кудинов В.А. Разработка системы идентификации речевого сигнала // Auditorium. 2019. № 1 (21). С. 72-79.

3. Иванов А.В., Рева И.Л., Шемшетдинова Э.Э. Исследование влияния различий в спектрах речи на результат оценки разборчивости // Динамика систем, механизмов и машин. 2017. Т. 5. № 4. С. 65-70.

4. Молдахмедова З.К., Ещанова Р.М., Молдахмедова Ж.К. Компьютерные технологии как эффективное педагогическое средство в обучении английскому языку // Теория и практика современной науки. 2018. № 12 (42). С. 291-293.

5. Зимин И.Ш. Методы обучения иностранному языку с использованием веб-приложения. // Научный аспект. 2020. Т. 4. № 3. С. 481-489.

6. Пыжьянова Е.А., Попова К.И., Лялин С.Г. Разработка и исследование алгоритма Voice Activity Detection // Общество. Наука. Инновации (НПК-2018). Сборник статей XVIII Всероссийской научно-практической конференции: в 3 томах. Вятский государственный университет. 2018. С. 853-859.

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 надежность искусственных нейронных сетей на базе мемристоров

Обеспечение надежности искусственных нейронных сетей на базе мемристоров (ИНСМ) и их компонентов предполагает проведение комплекса действий, повышающих показатели надежности до заданного уровня и (или) сохраняющих их на заданном уровне. Для решения этой задачи необходимо ослабить, устранить факторы, дестабилизирующие работу ИНСМ или компенсировать их деструктивное влияние на надежность.

На основе анализа научно-технических источников в области общей теории надежности [1] и надежности информационных систем [2], искусственных нейронных сетей [3], ИНСМ [4], теории и практики системной инженерии [5], теории ИНСМ [6], результатов имитационного моделирования [7], предложен уточненный (системный) вариант классификации основных факторов, снижающих надежность ИНСМ [8].

Системная классификация основных факторов, снижающих надежность ИНСМ:

- надсистемные, системные, подсистемные;
- объективные и субъективные;
- физические, информационные, физическо-информационные;
- тип платформы технической реализации;
- тип ИНСМ;
- вид структурно-функциональной декомпозиции;
- структурно-функциональный уровень исследования;
- варианты схмотехнических, конструктивных, технологических реализаций;
- погрешности проектирования, производства, эксплуатации ИНСМ;
- чувствительность к внутренним и внешним воздействиям;
- погрешности и ошибки информационного, математического, методологического, аппаратного и программного обеспечений;
- этап и режим функционирования;
- тип выполняемой функции;
- аппаратные и программные;
- цифровые и аналоговые;
- электрические режимы элементов схмотехнической реализации;
- низкая стабильность электрофизических параметров мемристивных устройств;
- высокая скорость деградации свойств мемристивных устройств при функционировании.

Дополнительная классификация факторов:

- с вероятностными и детерминированными параметрами;
- одномерные и многомерные;
- аддитивные и мультипликативные;
- статические и динамические;
- периодические и непериодические;
- линейные и нелинейные;
- зависимые и независимые.

Уточненный (системный) вариант классификации факторов, снижающих надежность ИНСМ позволяет более эффективно применять известные и разрабатывать новые методы и алгоритмы их устранения, ослабления, подавления и (или) компенсации результатов деструктивного влияния на надежность [9].

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

Литература

1. Gnedenko B.V. and I.A. Ushakov (1995) Probabilistic Reliability Engineering. John Wiley & Sons, NY.
2. Шубинский И.Б. Надежные отказоустойчивые информационные системы. Методы синтеза /И.Б. Шубинский. М.: «Журнал Надежность», 2016. 546 с.
3. Потапов И.В. Модели, методы и задачи прикладной теории надежности нейромкомпьютерных систем. Диссертация на соискание ученой степени доктора технических наук. Новосибирск, 2010. 348 с.
4. Meiran Zhao et al. Reliability of analog resistive switching memory for neuromorphic computing // Applied Physics Reviews. 2020. № 7.
5. Косяков А. Системная инженерия. Принципы и практика. Пер. с англ. / под ред. В. К. Батоврина. М.: ДМК Пресс, 2014. 624 с.
6. Галушкин А. И. Нейронные сети: основы теории / А. И. Галушкин. Москва: Горячая линия-Телеком, 2013. 496 с.
7. Борданов И.А., Щаников С.А. Перспектива применения имитационного моделирования при проектировании искусственных нейронных сетей на базе мемристоров // Алгоритмы, методы и системы обработки данных. 2019, № 2(40). С. 13-19.
8. Данилин С.Н., Щаников С.А., Борданов И.А., Зуев А.Д. Классификация основных факторов, снижающих отказоустойчивость и надежность искусственных нейронных сетей на базе мемристоров // Наука и образование в развитии промышленной, социальной и экономической сфер регионов России. XII Всероссийские научные Зворыкинские чтения: сб. тез. докл. Всероссийской межвузовской научной конференции. Муром, 7 февр. 2020г. – Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. С. 110-111.
9. Данилин С.Н., Щаников С.А., Борданов И.А., Зуев А.Д. Основные направления повышения отказоустойчивости и надежности искусственных нейронных сетей на базе мемристоров // Наука и образование в развитии промышленной, социальной и экономической сфер регионов России. XII Всероссийские научные Зворыкинские чтения: сб. тез. докл. Всероссийской межвузовской научной конференции. Муром, 7 февр. 2020г. – Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. С. 112-113.

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*

Основная причина проблем оптимизации искусственных нейронных сетей на базе мемристоров

Определение понятия «оптимизация» является общепринятым в научно-техническом сообществе. Оптимизация — это выбор наилучшего (оптимального) варианта цели и (или) процесса ее достижения. Наиболее надежным методом нахождения наилучшего варианта является сравнительная оценка всех возможных вариантов (альтернатив) по одному и (или) нескольким критериям [1]. Однако на практике это осуществимо в исключительных случаях, и вместо простого перебора вариантов применяется большая группа общих подходов, математических методов, алгоритмов и имитационного моделирования объектов и процессов [2,3].

Проведенный авторами обзор опубликованных работ показал, что основные положения системной инженерии (СИ) в них учитываются и применяются только фрагментарно [4]. Данный подход не соответствует принципу системности, согласно которому все известные, и не известные объекты, и процессы, явления или события в мире живой и не живой природы, человеческом обществе, представляют собой системы различной степени целостности и сложности [1]. (Под «целостностью» понимается принципиальная несводимость свойств системы к сумме свойств составляющих её физических, информационных, структурных, функциональных компонентов и не выводимость из последних свойств целого [5,6]).

Авторы доклада предложили общий подход к разработке методов оптимизации искусственных нейронных сетей на базе мемристоров (ИНСМ) на всех этапах их жизненного цикла на основе методологии СИ. Разработали системный вариант методов оптимизации на основе комплексного применения основных положений СИ в области ИНСМ. Приводится пример сравнения эффективности системного и несистемного подхода к поиску оптимальных решений для построения классического и нейросетевого аппроксиматора функций повышенной отказоустойчивости [7].

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

Литература

1. Новая философская энциклопедия. Т. 3 / под общ. ред. В.С. Степин и др. М.: Мысль, 2010. 692 с.
2. Борданов И.А., Щаников С.А. Перспектива применения имитационного моделирования при проектировании искусственных нейронных сетей на базе мемристоров // Алгоритмы, методы и системы обработки данных. 2019. № 2(40). С. 13-19.
3. Реклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике. — М.: Мир, 1986. — 400 с.
4. Косяков А. Системная инженерия. Принципы и практика. Пер. с англ. / под ред. В. К. Батоврина. М.: ДМК Пресс, 2014. 624 с.
5. Кулик С.Д. Элементы системного анализа для студентов старших курсов университета. Естественные и технические науки. 2018. № 11 (125). С. 373–377.
6. Данилин С. Н., Щаников С. А., Борданов И. А., Зуев А.Д., Пантелеев С.В. Актуальные задачи теории и практики системной инженерии // Радиотехнические и телекоммуникационные системы. 2020. №4. С. 42-55.
7. Данилин С.Н., Щаников С.А., Борданов И.А., Зуев А.Д. Основные направления повышения отказоустойчивости и надежности искусственных нейронных сетей на базе мемристоров // Наука и образование в развитии промышленной, социальной и экономической сфер регионов России. XII Всероссийские научные Зворыкинские чтения: сб. тез. докл. Всероссийской межвузовской научной конференции. Муром, 7 февр. 2020г. – Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. С. 112-113.

Еремеев С.В., Абакумов А.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

Классификация геообъектов на основе топологического анализа

На основе методов топологического анализа [1,2] разработан программный комплекс, позволяющий выполнять поиск необходимого природного объекта на растровом снимке для его дальнейшей классификации и обработки. Структура программного комплекса включает несколько подсистем: подсистему выделения областей интереса на снимке, подсистему построения баркодов, подсистему поиска схожих объектов, а также подсистему вывода и экспорта найденных объектов.

Подробно описан принцип выделения объектов интереса на снимках, построения баркодов и их сравнения. Для каждого выделенного на геоснимке пространственного объекта вычисляются топологические характеристики в виде чисел Бетти, которые являются основой для построения баркода [3,4]. Показан процесс разложения изображения на последовательность бинарных изображений для выявления устойчивых топологических характеристик. Продемонстрирован принцип сравнения баркодов для определения схожести выделенных областей интереса с эталонными объектами.

Приведены примеры использования программного комплекса для задачи поиска айсбергов на растровом изображении. Показаны результаты найденных объектов с разной степенью схожести относительно эталонов в зависимости от заданных параметров. Программный комплекс может быть использован для широкого спектра задач при анализе природных объектов на геоснимках, включая обработку данных за разное время и на разных масштабах.

В качестве примера показан процесс выделения айсбергов на рисунке 1(а). Фон на нём неоднородный (присутствует часть материка), из-за чего использовать простую функцию выделения не получится. На рисунке 1(б) показан результат выделения с использованием сравнения топологических характеристик шаблона и объектов интереса.

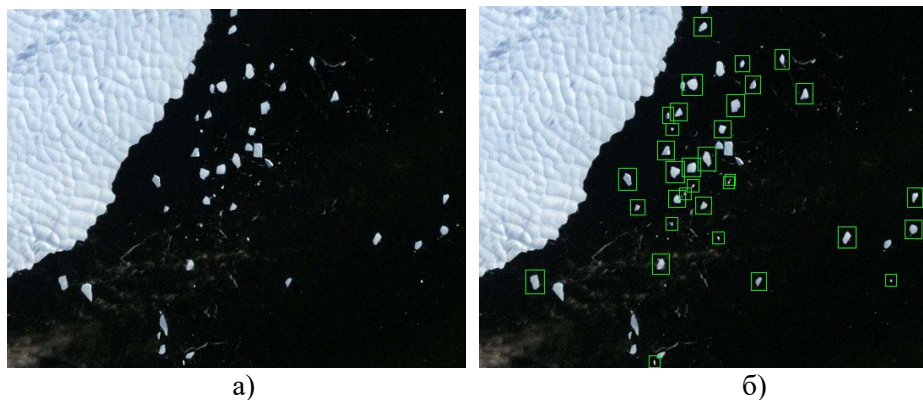


Рис. 1. Снимок (1999 г.) части дрейфующего в тёплых водах гигантского айсберга B10A и отделившихся от него айсбергов (а). Результат выделения объектов (б)

Литература

1. Еремеев С.В., Романов С.А. Алгоритм сегментации изображений на основе персистентной гомологии для решения задач поиска дефектов // Известия Юго-Западного государственного университета. 2020. Т. 24, № 1. С. 144-158.
2. Еремеев С.В., Минжилий Д.О. Исследование работы алгоритмов топологического анализа данных // Алгоритмы, методы и системы обработки данных. 2019. №1. С. 67-72.
3. Еремеев С.В., Ковалев Ю.А. Анализ временных эволюций объектов на основе компьютерной топологии // Оптико-электронные приборы и устройства в системах

распознавания образов, обработки изображений и символьной информации. Распознавание – 2018. Курск, 25-28 сентября 2018 г. С. 139-140.

4. Еремеев С.В., Купцов К.В., Ковалев Ю.А. Исследование алгоритма классификации пространственной информации на основе методов персистентной гомологии и random forest // IV международная конференция и молодёжная школа "Информационные технологии и нанотехнологии". Самара: Самарский национальный исследовательский университет имени академика С.П. Королева, 24-27 апреля 2018 г. С. 2384-2390.

Комкова С.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
savicheva.svetlana@gmail.com*

Обзор систем обработки изображений сетчатки глаза человека

Данная статья посвящена обзору современных медицинских систем, которые бы позволили производить всесторонний анализ изображений сетчатки глаза человека. Проведенное исследование позволяет утверждать, что в настоящее время нет подобных систем, которые бы отвечали предъявляемым требованиям. Поэтому актуальным является вопрос разработки системы, которая бы позволила выполнять операции обработки снимков сетчатки глаза с целью последующей диагностики по ним различных заболеваний глаза.

В настоящее время на рынке программного обеспечения представлен ряд программных продуктов, предназначенных для обработки офтальмологических снимков. Наиболее известные из них представлены ниже:

1. Универсальный АРМ получения и обработки медицинских изображений Medical Vision

Данная система предназначена для работы с цветными и черно-белыми медицинскими изображениями. Сюда относятся снимки, полученные с эндоскопа, УЗИ-снимки, кольпоскопия, офтальмология. Система позволяет получить снимок, произвести его цифровую обработку и сохранить вместе со всеми имеющимися данными о пациенте. При необходимости данные могут быть выведены на принтер.

2. Медицинская информационная система МЕДИАЛОГ

Аналогично предыдущей системе, данная система позволяет работать с различными медицинскими изображениями (рентгеновские снимки, результаты ангиографического обследования, офтальмологические снимки). Данные медицинские изображения можно подвергать обработке и редактированию. Врач при необходимости может оставлять комментарии, либо пометки непосредственно на снимках. Данные пометки сохраняются вместе со снимками в медицинской карте пациента, которую имеется возможность распечатать.

3. Программное обеспечение Mугіan

Это мультимодальное решение для просмотра и постобработки медицинских изображений. Включает простые в использовании, но мощные специализированные пакеты для колоноскопии, печени, легких, сосудов, ортопедии и других. Пользователь может получить комплект программ и клинических приложений в зависимости от специфики выполняемых исследований. С программой "Mугіan Expert" трёхмерное изображение становится необходимым инструментом, чтобы завершить диагностику, выявить патологию или поместить пораженный участок в его анатомический контекст.

Рассмотренные программные продукты представляют достаточно широкий спектр функций, связанных с обработкой медицинских изображений. Однако эти системы предназначены для обработки не только снимков сетчатки глаза, но и любых других медицинских изображений, что сказывается на их стоимости. При этом медицинскими учреждениями, специализирующимися на лечении глазных заболеваний, может быть не востребован весь функционал этих систем. Поэтому актуальной является задача разработки узкоспециализированного программного обеспечения, позволяющего обрабатывать исключительно снимки сетчатки глаза.

Литература

1. Комкова С.В. Предварительная обработка изображений сетчатки глаза // Естественные и технические науки. 2016, №6(96), С. 144-146.
2. Комкова С.В. Методика маркировки кровеносных сосудов на изображениях сетчатки глаза человека// Естественные и технические науки. 2019, №10, С. 250-252.

Подгорнова Ю.А.

*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*

Использование вейвлет-сегментации для выделения подозрительных областей на маммограммах

Маммография [1] – это неинвазивный метод выявления патологий молочных желез. Благодаря маммографическому скринингу можно выявить злокачественные новообразования молочных желез на ранних стадиях, тем самым дать шанс пациенту на эффективное лечение и полное выздоровление.

Маммограммы представляют собой малококонтрастные рентгеновские изображения молочных желез, состоящие из фона, области молочной железы и области грудной мышцы. Анализ маммографических снимков выполняется врачом рентгенологом визуально. Правильность постановки диагноза зависит от опыта и квалификации врача, а также от качества самого снимка.

Вейвлет-анализ получил применение в различных областях, в том числе и в обработке и анализе медицинских изображений.

Алгоритм сегментации маммограмм с использованием вейвлетов состоит из последовательности шагов, перечисленных ниже.

1) Предварительная обработка: маммографический снимок состоит из трех областей – фон, область молочной железы и область грудной мышцы. Для сегментации необходимо четко выделить область молочной железы, таким образом, сначала требуется бинаризовать изображение по методу Оцу [2], затем удалить лишние шумовые области на границе молочной железы и фона с помощью морфологической операции – эрозия [2].

2) Сегментация с использованием вейвлетов [3, 4, 5, 6]: изображение раскладывается на 3 вейвлета и масштабную функцию по базису Коэна – Добеши – Фово. В ходе экспериментов лучший результат был получен при разложении изображения на 2 уровня.

3) Постобработка: после вейвлет-преобразования выполняются операции постобработки для нахождения конечного сегментированного результата, сохраняя большие связанные бинарные объекты и соединяя соседние бинарные области. Так как обрабатываемое изображение на этом этапе может содержать шумовые области из-за разницы интенсивности областей новообразования и молочной железы, поэтому проводятся морфологические операции по заполнению областей и удалению любых лишних элементов, кроме новообразования. Области, относящиеся к темным углам вокруг изображения, удаляются бинарной маской на этапе предварительной обработки. Однако небольшие изолированные области сохраняются и соединяются вместе, если они находятся очень близко к новообразованию. С другой стороны, области, удаленные от опухоли, удаляются морфологическими операциями эрозией и дилатацией. Заключительным шагом является сглаживание неровных границ новообразования с помощью усредняющего фильтра [2].

Алгоритм был протестирован на 40 маммографических снимках из базы данных MIAS [2]. Тестирование показало, что с помощью данного алгоритма можно правильно выделить только новообразования на фоне жировой инволюции. Вероятность правильной работы алгоритма составляет 65.3%.

Литература

1 Высоцкая И. В. Клинические рекомендации Российского общества онкомаммологов по профилактике рака молочной железы, дифференциальной диагностике, лечению предопухолевых и доброкачественных заболеваний молочных желез / И. В. Высоцкая, В. П. Летягин, В. Г. Черенков и др. // Опухоли женской репродуктивной системы. 2016. Т. 12, № 3. С. 43-52.

2 Садыков С.С. Автоматизированная обработка и анализ маммографических снимков: монография/ С.С. Садыков, Ю.А. Буланова, Е.А. Захарова; Владим. гос. Ун-т им. А.Г. и Н.Г. Столетовых,- Владимир: Изд-во ВлГУ, 2014. - 208 с. ISBN 978-5-9984-0474-0.

3 Подгорнова Ю.А., Жизняков А.Л., Садыков С.С. Повышение контраста маммограмм, содержащих области рака молочной железы, на фоне жировой инволюции // Прикаспийский журнал: управление и высокие технологии. 2020. № 2 (50). С. 136-147.

4 Малла С. Вейвлеты в обработке сигналов. - М.: Мир, 2005. –671 с.

5 Блаттер К. Вейвлет-анализ. Основы теории. – М.: Техносфера, 2004. – 273 с.

6 Shehzad Khalid, Uzma Jamil, Kashif Saleem, M. Usman Akram, Waleed Manzoor, Waqas Ahmed, Amina Sohail. Segmentation of skin lesion using Cohen–Daubechies–Feauveau biorthogonal wavelet. Khalid et al. SpringerPlus (2016) 5:1603. DOI 10.1186/s40064-016-3211-4

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 синонимизированных заимствованных текстов с помощью «тяжелых» синонимов*

Одним из способов сокрытия заимствований текстов в настоящее время является синонимизация, то есть замена слов в тексте синонимами (словами со схожим смыслом, но различным написанием) [1]. При этом текст меняется таким образом, что системы проверки считают оценивают его как уникальный и не определяют наличие заимствований [2]. Синонимизация представляет собой лексическую замену некоторых слов текста их синонимами. Такая замена проводится в ручном (перефразирование) и автоматическом режиме с использованием так называемых синонимайзеров (онлайн сервисов или программ). В настоящее время существует достаточно большое количество синонимайзеров – Raskruty.ru, Usyn.ru, Seogenerator.ru, Seo-builder.ru, Sinoni.men, Sinonimov.ru, Rustxt.ru, Textrobot.ru, Online-sinonim.ru, Synonymizer.ru, Progaonline.com/synonymizer/, Fromtlt.ru/sinonim и т.д. Они в значительной степени отличаются используемыми базами синонимов и алгоритмами работы (в первую очередь степенью переработки текста и его «читаемостью») [3].

В настоящее время проблему обнаружения и обработки синонимизации предлагается решать путем расширения запросов сравнения синонимами (например, с использованием WordNet) [4, 5, 6]. Другим решением является сравнение словарей (наборов слов) проверяемых текстов, расширенных с помощью синонимов и очищенных от часто употребляемых слов [7, 8]. Подобные решения имеют один существенный недостаток – они работают с текстом как набором слов и не позволяют учитывать их взаимное расположение. По этой причине возникают проблемы с визуализацией найденных совпадений и корректного подсчета заимствований.

Проблему можно решить с помощью изложенного ниже подхода с использованием «тяжелых» синонимов. Для этих целей будет использоваться три словаря:

1. Словарь термов (слов) коллекции документов W. Для этих целей используется словарь системы Автор.NET (более 160 тысяч слов) [9].
2. Частотный словарь русского языка F [10].
3. Словарь синонимов S (база синонимов SynMaster).

Рассмотрим процедуру определения «тяжелых» синонимов:

1. Для каждого слова w_i в словаре W поставим в соответствие его вес f_i , подсчитанный на основе глобальной частоты его встречаемости в русскоязычных текстах. При отсутствии значения в таблице глобальных частот вес принимается равным 0.
2. Для каждого слова w_i по словарю S составляется список синонимов $w_i = \{s_1, s_2 \dots s_{in}\}$.
3. Для каждого синонима подсчитываются их веса $f_i = \{f_{i1}, f_{i2} \dots f_{in}\}$.
4. Определяется наиболее представительный («тяжелый») синоним в соответствии с максимальным значением веса, т.е. $\max(f_{i1}, f_{i2} \dots f_{in}) \rightarrow s_i$.
5. Если вес синонима s_i превышает вес слова w_i , то синоним принимается как кандидат на замену слова, в противном случае процедура поиска синонимов прекращается (переходим на шаг 7).
6. Синоним-кандидат s_i считается как исходное слово w'_i и действие итерационно повторяется для него (переходим к шагу 2).
7. Слово w_i заменяется найденным синонимом s_i . При весе слова больше весов всех кандидатов на синонимы, его замена синонимами не производится, а слово считается наиболее представительным.

Как можно заметить, процедура поиска синонимов заключается в выявлении синонимов с максимальным весом, затем для найденных синонимов также ищется синоним с максимальным

весом и так далее, до тех пор, пока не будет найден синоним, вес у которого больше всех остальных кандидатов другими словами самый «тяжелый» синоним.

Основная идея подхода заключается в том, что какой бы из синонимов не был выбран в момент синонимизации (и из какой базы синонимов), он вероятнее всего попадет в рассматриваемый список синонимов. Так как мы не знаем, какое именно слово было заменено на какой синоним, мы можем заменить все слова ни их самые вероятные («тяжелые») синонимы. В этом случае и исходный и проверяемый (прошедший синонимизацию для сокрытия заимствований) текст будут изменены примерно одинаково на одни и те же «тяжелые» синонимы. То есть не нарушается структура текста и расположение входящих в него слов (сами слова могут при этом быть заменены). Поэтому сравнение их содержания даст вполне приемлемые результаты.

Исследования показали, что часто в качестве синонимов могут встречаться местоимения и другие часто употребляемые слова. Рассмотрим, например, список синонимов для слова «автор»:

Синоним	Вес
я	15631.11
мы	4740.29
писатель	213.5
автор	160.56
композитор	17.93
литератор	16.04
создатель	13.04
творец	12.36
виновник	9.3
сочинитель	4.71
составитель	2.2
полиграф	2.2
либреттист	0
компилятор	0
доксограф	0
оригинатор	0
ткомедиограф	0
песенник	0
комедиограф	0
авторикша	0

Как можно заметить, в списке встречаются местоимения «я», «мы», занимающие 6 и 23 место по частоте встречаемости в русском языке. Естественно, появление таких частотных слов приводит к признанию их лучшими кандидатами в синонимы, а сам текст может превратиться в набор частотных слов. Для того, чтобы избежать этого, следует проводить фильтрацию кандидатов в синонимы, удаляя наиболее частотные слова (так называемые стоп-слова). В нашем случае был выбран порог веса в 500, что обеспечивает удаление 190 наиболее частотных слов русского языка. Таким образом, «тяжелым» синонимом слова «автор» будет слово «писатель».

Литература

1. Шабанова С.А. Сущность явлений синонимии и синонимизации // Новый университет. Серия: Актуальные проблемы гуманитарных и общественных наук, № 9, 2012, С. 48-50.
2. Чиркин Е.С. Системы автоматизированной проверки на неправомерные заимствования // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2013. № 12 (128). С. 164-174.
3. Шарапова Е.В. Текстовые заимствования и борьба с ними: монография / Е. В. Шарапова; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2020. – 160 с.
4. Chong, M., and Specia, L. 2011. Lexical Generalisation for Word-level Matching in Plagiarism Detection. In RANLP, pp. 704-709.

5. Nawab, R. M. A., Stevenson, M., and Clough, P. 2016. An ir-based approach utilising query expansion for plagiarism detection in medline. IEEE/ACM transactions on computational biology and bioinformatics.
6. Nawab, R. M. A., Stevenson, M., and Clough, P. 2012, April. Retrieving candidate plagiarised documents using query expansion. In European Conference on Information Retrieval. Springer Berlin Heidelberg, pp. 207-218
7. Atadjanov J. Document Plagiarism Detection On The Internet With Accounting Synonym Forms Of Words // International journal of scientific and technology research, vol. 8, issue 12, 2019, 1517-1519.
8. Зиберт А.О., Мирошниченко В.В. Об использовании словарей синонимов в алгоритме определения наличия заимствований в тексте // Universum: технические науки. 2014. № 12 (13). С. 3
9. Sharapova E. One way to fuzzy duplicates detection // International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM. 14. 2014. С. 273-278.
10. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). - М.: Азбуковник, 2009. <http://dict.ruslang.ru/freq.php>

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

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 распределенной системы поиска текстовых заимствований *

Исследования показали, что для обеспечения достаточной производительности системы проверки оригинальности текстов необходимо использовать распределенную модульную структура [1, 2, 3]. При этом каждый модуль призван выполнять определенный круг задач и в значительной степени быть независимым от других.

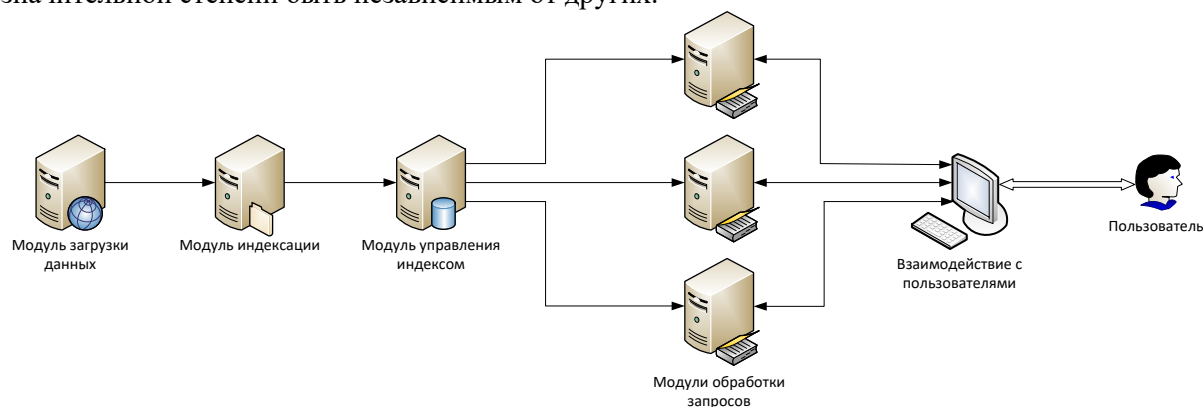


Рис. 1. Структура распределенной системы поиска заимствований.

1. Модуль загрузки данных

Основное назначение модуля – сбор данных (документов) для наполнения поискового индекса. Основным источником данных является сети Интернет. Модуль выполняет следующие действия:

- Загружает документы с вебсайтов
- Преобразует их в машиночитаемый формат (в нашем случае в обычный текст без разметки)
- При необходимости производит перевод текста в единую кодировку (UTF-8)
- Производит очистку полученного текста от элементов разметки, спецсимволов, выполняет слияние слов с переносами, разорванных переходами на новую страницу предложений

2. Модуль индексации

Модуль получает загруженные документы и производит их индексацию. При этом осуществляется подсчет сигнатур, вычисление TF*IDF индекса, набора шинглов и т.д. Модуль содержит средства нормализации входящих в документ слов, фильтрации стоп-слов (наиболее часто встречающихся слов во всей коллекции документов).

Модуль позволяет искать уже имеющиеся копии добавляемого документа в поисковом индексе, а также обновлять поисковый индекс более свежими версиями документов.

3. Модуль управления индексом

Пополнение поискового индекса в реальном времени чаще всего представляет собой сложную задачу, требующую больших вычислительных затрат, что оказывает влияние на производительность всей системы. В первую очередь это связано с необходимостью перестройки упорядоченных индексов, размеры которых могут составлять сотни гигабайт [4].

По этой причине чаще всего поисковый индекс обновляется пакетно. Иногда, для ускорения обновления индекс разбивается на две части – основной статический (большой) индекс, и дополнительный динамический (обновляемый) индекс, в который записываются все вновь добавленные документы. При очередном цикле обновления данные из дополнительного индекса переносятся в основной индекс, а дополнительный индекс очищается в ожидании новых данных.

Всеми операциями создания и модификации основного и дополнительного индексов занимается модуль управления индексом. Он также производит проверку индексов, удаления из них дублирующихся данных, производит подсчет статистики и т.д. Итогом работы модуля являются готовые индексные файлы, передаваемые системе обработки запросов.

4. Модули обработки запросов

Для повышения производительности системы предлагается вместо единого хранилища индексов использовать распределенное хранение. При этом, в зависимости от нагрузки к системе можно производить работу с поисковым индексом либо целиком в каждом модуле, либо разбив его на части (каждый модуль работает со своей частью индекса – шинглами, сигнатурами, $TF*IDF$ и т.д.) [5].

Число модулей обработки запросов зависит от количества обращений к системе проверки. Это позволяет добиться масштабируемости системы без снижения ее производительности.

5. Интерфейс взаимодействия с пользователями

Интерфейс взаимодействия с пользователями предполагает получение документов для проверки от пользователей, их предварительную обработку, отправку запросов на поиск похожих документов модулям обработки запросов, получение результатов их работы и выдачу отчетов пользователям.

Литература

1. Sharapova E.V. Computational load reduction of fuzzy duplicate detection in large amounts of information // IOP Conference Series: Materials Science and Engineering, vol.734, 2020. 012119.
2. Sharapova E. Increase the Speed of Search Index in the Duplicate Text Detection Systems // 2020 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Svetlogorsk, Russia, 2020, pp. 1-5.
3. Sharapova E.V. One way to solve the problem of fuzzy text duplicates detection // Materials of the International Conference “Scientific research of the SCO countries: synergy and integration” - Reports in English (July 12, 2019). Beijing, PRC, 2019. P.59-64.
4. Шарапова Е.В. Структура поискового индекса системы обнаружения нечетких дубликатов текстов // Наука и образование в развитии промышленной, социальной и экономической сфер регионов России. XII Всероссийские научные Зворыкинские чтения: сб. тез. докл. Всероссийской межвузовской научной конференции. Муром, 7 февр. 2020 г. – Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. – [Электронный ресурс]: 1 электрон. опт. диск (CD-ROM). С128-129.
5. Шарапова Е.В. Обнаружение нечетких дубликатов текстов в больших массивах информации с помощью сигнатур содержания // Управление развитием крупномасштабных систем MLSD'2019 Материалы двенадцатой международной конференции (1–3 октября 2019 г., Москва). – М. Институт проблем управления им. В.А. Трапезникова, 2019. С. 1009-1011.

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