СРАВНЕНИЕ АЛГОРИТМОВ ВЫЧИСЛЕНИЯ РЕДАКЦИОННОГО РАССТОЯНИЯ НА ПРИМЕРЕ МЕДИЦИНСКИХ ЗАКЛЮЧЕНИЙ

COMPARISON OF ALGORITHMS FOR COMPUTING THE DRAFTING DISTANCE ON THE EXAMPLE OF MEDICAL REPORTS

V. Yurkin
I. Radchenko
A. Yarkin

Summary. This article describes the algorithms for finding the editorial distance, their evaluation based on their application to solve the problem of extracting the data necessary for the study (the name and initials of the patient, the number and date of the study and others), from the medical reports of the known format. The article describes and uses such algorithms as: distance of Levenshtein, similarity of Jaro-Winkler, distance of Damerau-Levenshtein, Hamming distance.

Keywords: algorithms of editorial distance, computer modeling, mathematical modeling, medical conclusions.

Юркин Вадим Михайлович
Аспирант, Санкт-Петербургский НИИ Информационных Технологий Механики и Оптики
vuy-ifmo@gmail.com
Радченко Ирина Алексеевна
К.т.н., доцент, Санкт-Петербургский НИИ Информационных Технологий Механики и Оптики
Яркин Антон Сергеевич
Аспирант, Санкт-Петербургский НИИ Информационных Технологий Механики и Оптики

Аннотация. В данной статье описаны алгоритмы поиска редакционного расстояния, дана их оценка на основе их применения для решения задачи извлечения данных, необходимых для исследования (фамилия и инициалы пациента, номер и дата исследования и других), из медицинских заключений известного формата. В статье описаны и использованы такие алгоритмы как: расстояние Левенштейна, сходство Джаро-Винклера, расстояние Дамерау-Левенштейна, расстояние Хэмминга.

Ключевые слова: Алгоритмы редакционного расстояния, компьютерное моделирование, математическое моделирование, медицинские заключения.

Введение

Современная медицина и здравоохранение за последние десятилетия поднялись на недосягаемый ранее уровень. Значительно улучшилась техническая оснащенность медицинских учреждений, появилась возможность диагностировать заболевание на самой ранней стадии, обеспечить быстрейшее выздоровление и восстановление работоспособности обратившегося за помощью человека.

Понятие медицина в первую очередь дает обозначение сфере человеческой деятельности, которая направлена на доскональное изучение и рассмотрение процессов, которые осуществляются в организме каждого человека. Цель и основополагающая задача очень очевидна и проста — это регулярное ведение различных разработок, предостережения, диагностики и последующего лечения человека от всевозможных болезней. Современная медицина ведет постоянные разработки инновационных технологий, которые предоставляют уникальные возможности заранее предупредить возникновение болезней.

Компьютерная томография на сегодняшний день — ведущий метод диагностики многих заболеваний головного мозга, позвоночника, легких и средостения, печени, почек, поджелудочной железы, надпочечников, аорты и легочной артерии, сердца и ряда других органов. Компьютерную томографию можно использовать и как метод первичной диагностики, и как уточняющую методику, когда предварительный диагноз уже поставлен с помощью УЗИ или клинического обследования.

На компьютерной томограмме [1] видны опухоли, камни, кисты. Таким образом, КТ является практически универсальным методом диагностики, позволяющим врачу увидеть максимально подробную картину состояния организма. Для повышения информативности КТ его выполняют с использованием контрастного вещества (в частности, при изучении сосудов и полых органов).

Диагностическая станция производит не один файл, а сразу несколько для одного исследования. Эти файлы имеют логическую структуру. Файлы объединяются в серии и представляют собой набор последовательных срезов какого-либо органа. Серии объединяются в стадии. Стадия определяет всё исследование. Последовательность серий в стадии определяется протоколом исследования.

Таким образом, после завершения работы компьютерного томографа на выходе врач получает исследование в формате DICOM [2] — набор снимков, объединенных идентификатором исследования. Далее при помощи специального программного обеспечения врач

проводит анализ полученных изображений и по итогам врачебного обследования составляется развернутое медицинское заключение, которое содержит оценку состояния здоровья пациента и вытекающие из этого рекомендации.

Врачебное заключение — это бланк, заполняемый в формате, определенном медицинским учреждением. В большинстве случаев форма медицинского заключения помимо диагноза и рекомендаций содержит и данные пациента, такие как фамилия, имя, отчество, дата рождения, а также информацию об исследовании.

К сожалению, чаще всего бланки медицинских заключений хранятся либо на бумажных носителях, либо в стандартных форматах документов: DOC, DOCX, PDF, ODT. Эти бланки просто заносятся в архив и не связаны непосредственно с исследованиями компьютерного томографа. Точно идентифицировать исследование по медицинскому заключению представляется возможным только лишь при помощи врача или же самого пациента. Кроме того, информация, написанная в заключении, может содержать опечатки и неточности, что еще больше может вводить в заблуждение. Так происходит, потому что в объединении КТ-исследований и врачебных заключений изначально не было нужды. Однако сейчас, в эпоху активного развития компьютерных технологий и различных интеллектуальных систем, такая связь оказывается очень полезной.

Основной задачей данной статьи и является исследование и разработка алгоритма, позволяющего находить и извлекать из врачебных заключений такие данные, как фамилия и инициалы пациента, номер и дата исследования, с учетом ошибок и опечаток производить идентификацию снимков компьютерной томографии. Идентификация подразумевает однозначное определение нужного DICOM-исследования компьютерного томографа по данным, извлекаемым из врачебного заключения.

Медицинское заключение является документом, который содержит основные данные клиента и медицинского учреждения, а также результаты исследования и рекомендации касательно лечения.

Формат медицинского заключения может кардинальным образом отличаться в зависимости от медицинского учреждения. Кроме того, в различным клиниках, для работы с заключениями используется разное программное обеспечение. Кто-то использует бумажные носители, кто-то предпочитает работать с документами Word, ODT и т.п.

В данной работе будут рассмотрены врачебные заключения одного из крупных медицинских исследовательских центров. Сами заключения хранятся в файлах DOCX. При создании нового заключения врачи используют готовый шаблон, куда вносят данные пациента, а также записывают дату, результат обследования и рекомендации. Однако, так как шаблон уже поставляется как DOCX-файл, врач волен изменять его. Таким образом каждое конкретное медицинское заключение может быть по-своему уникально. Различного рода переносы строк, создание новых абзацев, лишние пробелы. Это лишь краткий перечень того, каким образом заключения могут меняться от случая к случаю.

Врачебные заключения

Рассмотрим пример врачебного заключения (рис. 1). Здесь можно наблюдать следующую структуру. Вначале идёт параграф с названием медицинского учреждения, затем информация об отделении. Ниже расположена таблица, в которой находится информация о пациенте и о его обследовании:

- Фамилия, имя и отчество пациента, где имя и отчество пишутся как инициалы;
- Возраст пациента;
- Номер отделения, в котором проводилось обследование;
- Номер исследования, который проставляется врачом и не является номером исследования, который проставляется компьютерным томографом;
- Текст заключения, который впоследствии может быть проанализирован и использован для обучения, например, экспертной системы.

Под таблицей расположен параграф с фамилией и инициалами врача, проводившего обследование, и дата проведения обследования. При этом дата обследования может отличаться от даты исследования на компьютерном томографе.

Такой шаблон заключения используется всеми врачами медицинского учреждения, но, как уже было сказано, он может слегка меняться от случая к случаю даже по чистой случайности.

Использование алгоритмов нахождения редакционного расстояния для валидации данных

На данном этапе появляются две серьезные проблемы, которые кардинальным образом могут повлиять на результат идентификации. Такими проблемами являются, во-первых, опечатки в тексте, допущенные врачом при вводе фамилии или инициалов пациента. А во-вторых, важной деталью здесь является запись данных пациента на транслите. Когда врач вводит фа-

«Российский научно-исследовательский нейрохирургический институт им. проф. А.Л. Поленова» - филиал ФГБУ «СЗФМИЦ» МЗ РФ

Кабинет компьютерной томографии

Телефоны: зав отделением 273-76-62, компьютерный томограф 273-81-84

ФИО	Шухамжаева Х.Ю.	Возраст	4г
Номер отделения	5	Номер исследования	12345678

На МСК головного мозга смещения срединных структур нет. Водянки нет. Боковые желудочки асимметричны S>D. Выявляется пластинчатая острая субдуральная гематома правой лобно-височной области толщиной 4мм. Субарахноидальные пространства сужены в правой гемисфере. При исследовании в режиме «костного окна» костно-травматических изменений не выявлено.

Заключение: Пластинчатая острая субдуральная гематома правой лобно-височной области.

Врач: Потемкина Е.Г. Дата: 18.11.2017г.

Рис. 1. Элементы структуры файла формата DOCX на примере врачебного заключения

милию и инициалы пациента в аппарат компьютерной томографии, то делает он это именно буквами латинского алфавита, поскольку, как уже было сказано, следуя стандарту DICOM требуется использовать именно латинский алфавит и, следовательно, транслитерацию. В связи с этим, как уже было сказано ранее, запись различных букв русского алфавита может быть довольно произвольной.

Таким образом, появляются два источника ошибок и опечаток, которые встречаются весьма часто. Так одна и та же фамилия одного и того же пациента может быть записана по-разному и возможно даже иметь разную длину. Записанная с ошибками и опечатками фамилия пациента может также отчасти походить на фамилию другого пациента. Для того, чтобы различать такие фамилии и проводить корректную идентификацию исследований, было решено использовать один из алгоритмов вычисления редакционного расстояния. Правильно подобранный алгоритм вычисления редакционного расстояния позволяет с высокой точностью находить наиболее похожие строки.

В данной работе для расчета редакционного расстояния были выбраны уже существующие и наиболее популярные алгоритмы, которые используются именно при работе с текстом и подходят для решения поставленной задачи. Однако окончательное решение об использовании конкретного алгоритма можно принять лишь после получения результатов работы всех алгоритмов, их анализа и сравнения. Каждый из выбранных алгоритмов производит расчет минимального количества операций, необходимых для преобразования одной строки в другую. Таким образом, проводя вычисление редакционного расстояния для каждой пары фамилий (из текста врачебного заключения и из базы данных PACS) можно составить рейтинг наиболее схожих строк и тем самым допустить опечатки и ошибки, которые могли иметь место при записи данных человеком. Как уже говорилось ранее, компьютерный томограф хранит данные об исследованиях,

используя исключительно буквы латинского алфавита. В то же время, текст любого врачебного заключения использует только буквы русского алфавита. По этой причине существует два способа идентификации исследований по фамилии и инициалам. В первом случае можно перевести фамилию и инициалы пациента из врачебного заключения применив транслитерацию. В таком случае сравнение будет производиться с оригинальными данными исследования, находящимися в базе данных PACS. Во втором случае можно перевести фамилию и инициалы пациента из базы данных PACS на русский язык.

Рассмотрим сначала первый вариант перевода. Для этого необходимо сперва привести фамилию и инициалы из исследования и фамилию и инициалы из медицинского к одному формату: убрать лишние символы, привести строки к верхнему регистру. Затем производится перевод русских букв в ФИО из заключения на буквы латинского алфавита.

Далее происходит вычисление коэффициента схожести строк (нахождение значения редакционного расстояния) с использованием одного из вышеупомянутых алгоритмов. При этом стоит учесть тот факт, что редакционное расстояние всегда вычисляется лишь для фамилии пациента. Инициалы состоят из одной русской буквы (либо из нескольких букв при работе с латинским алфавитом) и несовпадение инициалов должно вести к неудачной идентификации, поскольку у двух разных людей может быть одна фамилия, но инициалы с большой долей вероятности будут отличаться.

Так, произведя вычисление редакционного расстояния и проверку инициалов, найденные исследования сортируются по убыванию коэффициента схожести. На первом месте в таком случае располагается исследование с наиболее похожей фамилией.

Проведя анализ результатов работы алгоритмов выяснилось, что при схожести строк более 90% можно утверждать, что две строки равны, но допускается небольшое расхождение: не хватает одной буквы, буквы поменяны местами, в строке содержится лишняя буква. Вид расхождения зависит от алгоритма.

Более того, в некоторых случаях строки можно считать равными и при меньшем проценте схожести, однако в таком случае могут встречаться и неверные исследования, поэтому брать слишком низкий порог схожести нельзя. Особенно такой пример актуален для коротких строк, где изменение одной буквы сильно влияет на результат проверки.

Также среди 100% идентифицированных исследований могут встречаться дубликаты. Это означает, что среди исследований за определенную дату было найдено несколько экземпляров со 100% схожестью. При этом такие исследования могут вообще не относиться друг к другу и иметь разные идентификаторы. В таких случаях нужно более внимательно изучать состав этих исследований и затем уже совмещать их с соответствующими заключениями.

Таким образом результаты работы каждого из алгоритмов можно отнести к одной из следующих групп:

1. Идентифицированные исследования — 100% схожесть строк без дубликатов;
2. Идентифицированные исследования с дубликатами — 100% схожесть строк и наличием одного или нескольких дубликатов;
3. Идентифицированные исследования (более 90% схожести) — исследования с более чем 90% схожестью считаются идентифицированными;
4. Неидентифицированные исследования — менее 90% схожести, такие исследования считаются неидентифицированными.

Используя такую группировку можно сравнить результаты работы представленных алгоритмов при использовании букв исключительно латинского алфавита (рис. 2). Из графика видно, что все алгоритмы одинаково хорошо показали себя при однозначной идентификации исследований, а также при однозначной идентификации с наличием дубликатов. Это не удивительно, поскольку однозначная идентификация предполагает полное сходство двух строк. Следовательно, все фамилии и инициалы пациентов, правильно переведенные с русских букв на буквы латинского алфавита, во всех случаях выдают стопроцентное сходство.

Расхождения в результатах начинаются, когда происходит неполное сходство ФИО из врачебных заключений и из исследований в базе данных PACS. Можно наблюдать, что алгоритм Дамерау-Левенштейна в таком случае работает лучше, хотя и всего на 3 позиции. Это обусловлено тем, что разница алгоритмов Левенштейна и Дамерау-Левенштейна состоит лишь в том, что алгоритм Дамерау-Левенштейна поддерживает операцию транспозиции символов, то есть перестановку. Такие ошибки, как видно из графика, встречаются крайне редко, но именно в таких случаях алгоритм Дамерау-Левенштейна является более эффективным.

Кроме того, стоит обратить внимание и на то, что алгоритм Джаро-Винклера, который часто используется именно для поиска опечаток и прочих ошибок в тексте, показал наилучший результат, а алгоритм Хэмминга наоборот, произвел наименьшее количество идентификаций исследований.

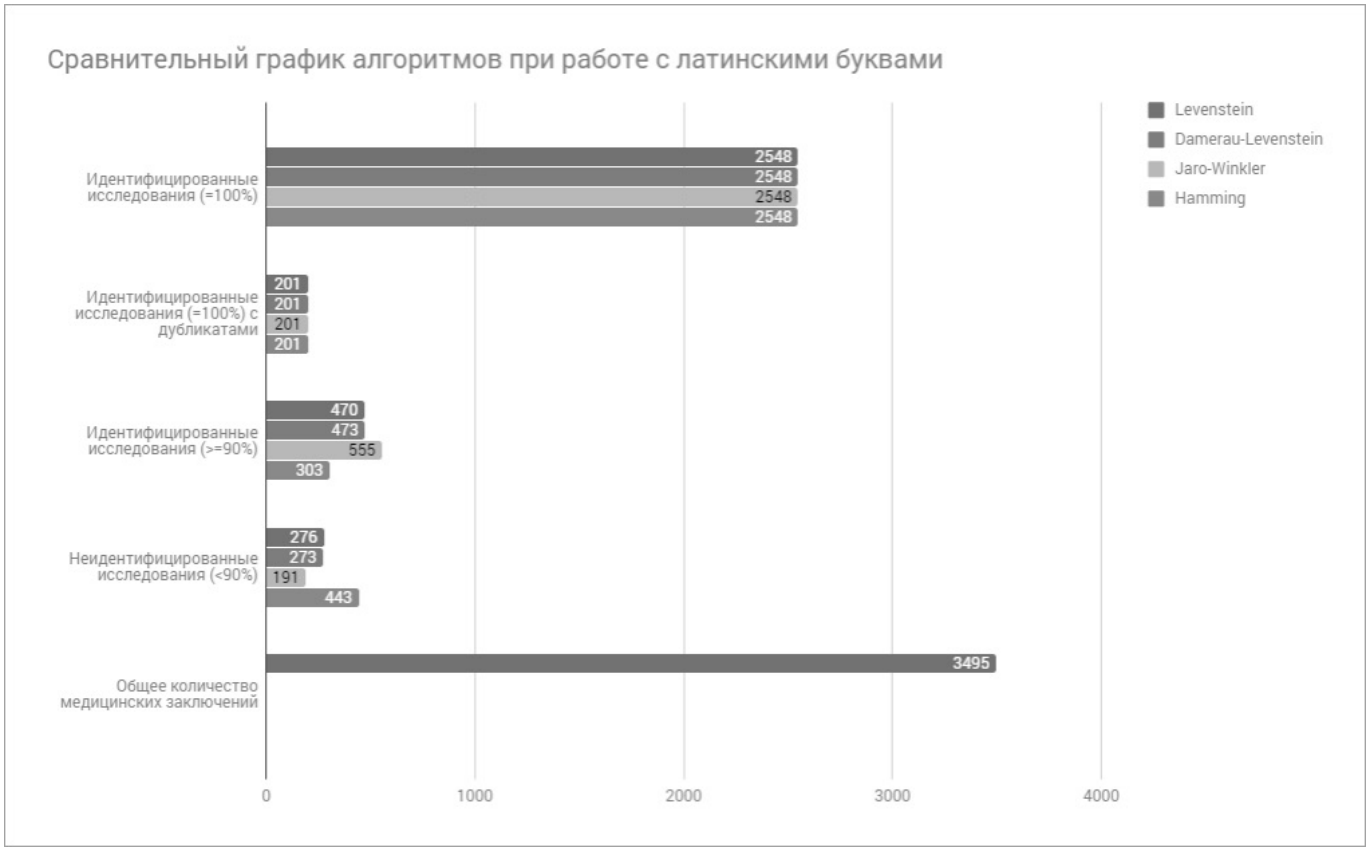

Рис. 2. Сравнительный график результатов работы алгоритмов при использовании букв латинского алфавита

Качество работы алгоритма Хэмминга объясняется его простотой. Более того данный алгоритм не применяется к строкам разной длины, а из-за возможных неточностей, вызванных переводом русских букв на латинские, такие случаи могут встречаться достаточно часто.

Попытка идентификации КТ-исследований с использованием оригинальных фамилий и инициалов из базы данных и с переводом данных медицинских заключений на латинские буквы выдала довольно хорошие результаты. Однако помимо данного способа необходимо также проверить и вариант с использованием исключительно букв русского алфавита.

При таком подходе фамилия и инициалы пациента из врачебного заключения остаются неизменными, они лишь приводятся к верхнему регистру. Преобразованию подвергается фамилия и инициалы, полученные из исследований, которые хранятся в базе данных PACS.

Преимуществом данного метода идентификации является то, что при переводе букв латинского алфавита на буквы русского алфавита появляется возможность однозначной замены латинской буквы или сочетания латинских букв на русскую букву. Это повышает корректность перевода и тем самым увеличивается степень схожести двух строк.

Для данного подхода к идентификации также имеет смысл провести анализ эффективности и времени работы, как это было сделано в случае с буквами латинского алфавита. На рисунке 15 изображен сравнительный график с результатами идентификации по каждому из алгоритмов.

Из графика следует, что при работе с буквами русского алфавита количество однозначно идентифицированных исследований выросло практически на 300 штук. Такой прирост свидетельствует об эффективности перевода латинских букв на русские, потому как результаты получаются более корректными и те ошибки, которые могли быть допущены при переводе букв русского алфавита на буквы латинского алфавита, в данном случае не были совершены.

Однако, если обратить внимание на количество идентифицированных исследований, чей показатель схожести строк меньше 100%, то можно заметить, что

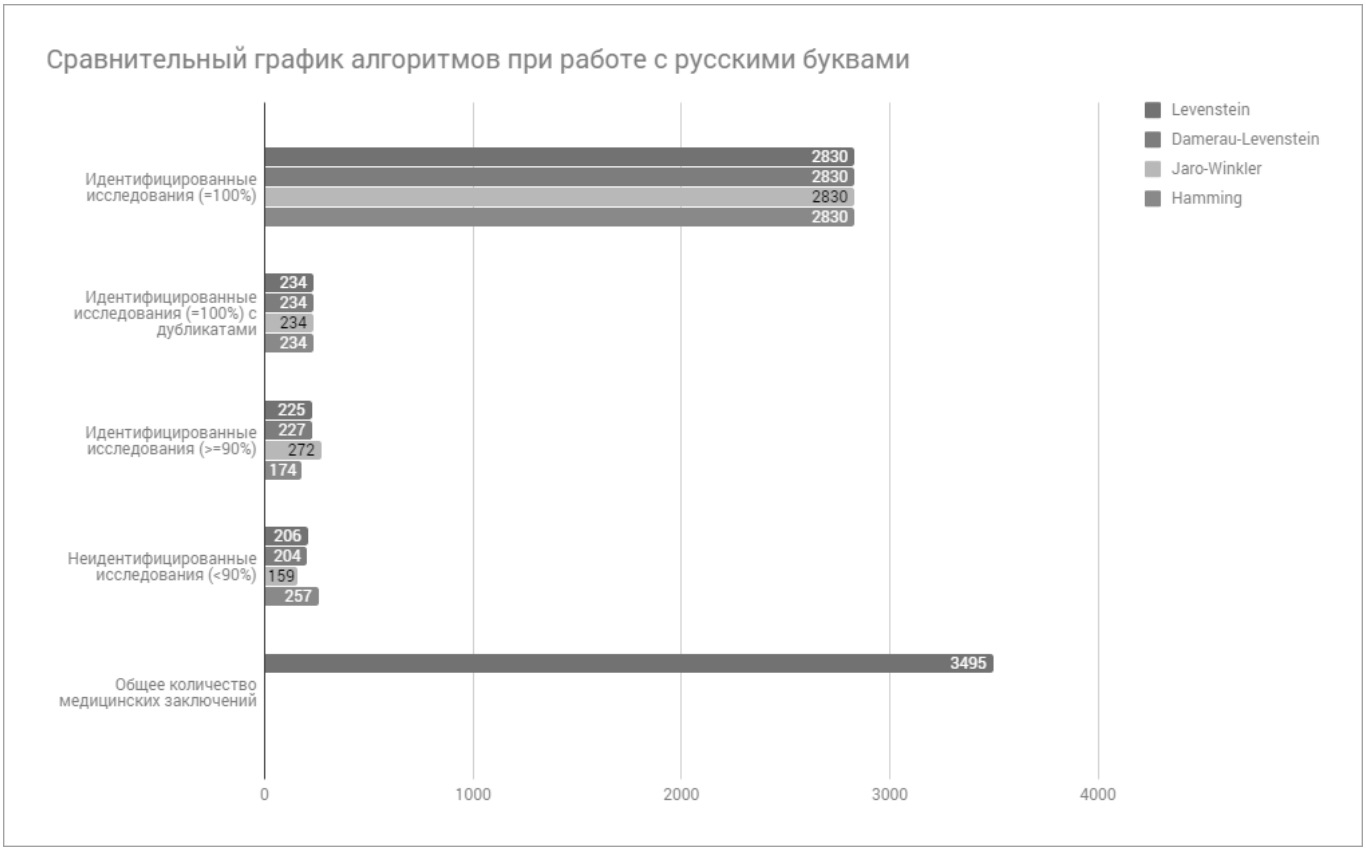

Рис. 3. Сравнительный график результатов работы алгоритмов при использовании букв русского алфавита

их количество наоборот уменьшилось. Это объясняется тем, что исследования, чья степень схожести составляла менее 100% при работе с буквами латинского алфавита, переместились в число однозначно идентифицированных исследований. Таким образом произошла своеобразная компенсация.

Как и в прошлый раз, использование алгоритма Джаро — Винклера позволило произвести наибольшее количество идентификаций, а алгоритм Хэмминнга показал себя хуже остальных в данной задаче. Алгоритмы Левенштейна и Дамерау — Левенштейна показали практически одинаковые результаты. Разница составила всего 2 исследования в пользу алгоритма Дамерау — Левенштейна.

Вывод

В результате был получен наиболее эффективный способ идентификации исследований, который включает в себя извлечение данных из документов врачебных заключений по структуре файла, перевод фамилии и инициалов пациента из базы данных PACS на буквы русского алфавита, использование алгоритма Джаро-Винклера для вычисления редакционного расстояния.

Разработанный алгоритм на данном этапе поддерживает работу с заключениями конкретного медицинского учреждения, однако он может быть использован и другими медицинскими учреждениями. Для этого необходимо добавить поддержку новой структуры документа, указав место расположения ФИО пациента и даты исследования.

Области применения данного алгоритма достаточно обширны. Связь текста заключения и снимков DICOM-исследования как минимум облегчит работу самим врачам, поскольку идентификатор исследования, хранящийся в базе данных PACS, может быть использован в целях объединения документа врачебного заключения с DICOM-исследованием. Стандарт DICOM позволяет конвертировать PDF-файлы в файлы DICOM и, следовательно, проставить необходимые теги в результирующем файле. Таким образом идентификатор исследования, полученный после идентификации из базы данных PACS, может быть записан в соответствующий тег DICOM-файла заключения. Затем этот файл может

быть внесен в систему PACS для дальнейшей обработки. В конце концов при поиске исследования врачом система PACS будет возвращать не только данные снимков исследования, но и PDF-файл заключения, который был соединен с исследованием.

Помимо использования разработанного алгоритма в целях помощи медицинским работникам, полученный метод идентификации может быть использован для извлечения набора связанных данных. Это означает, что по тексту врачебных заключений возможно проводить классификацию идентифицированных снимков и находить снимки, на которых были выявлены конкретные виды заболеваний. Кроме того, благодаря классификации изображений появляется возможность обучать различные экспертные системы в целях автоматической классификации изображений, либо при разработке систем помощи принятия решений.

ЛИТЕРАТУРА

1. Хофер, М. Компьютерная томография. Базовый курс / М. Хофер.— Медицинская литература, 2011–232 с.
2. Пьяных, О. Digital Imaging and Communications in Medicine (DICOM) О. С. Пьяных.— Springer-Verlag Berlin Heidelberg, 2008–384 с.
3. Christen, P. A Comparison of Personal Name Matching: Techniques and Practical Issues [Электронный ресурс] / P. Christen.— Электронные данные — The Australian National University, 2006 — Режим доступа: http://users.cecs.anu.edu.au/~Peter.Christen/publications/tr-cs-06–02.pdf
4. Peck, A. Clark's Essential PACS, RIS and Imaging Informatics /A. Peck.— CRC Press, 2017–248 с.
5. Wagner, R. A. The String-to-String Correction Problem / R. A. Wagner, M. J. Fischer — ACM New York, 1974.