D. M. Klionskiy
Saint Petersburg Electrotechnical University «LETI»

MULTIDIMENSIONAL ALGORITHM FOR PROCESSING VECTOR HYDROACOUSTIC SIGNALS AND ITS SOFTWARE IMPLEMENTATION

The paper discusses one-dimensional and multidimensional weighted overlap-add algorithm for processing one-dimensional and multidimensional (vector) hydroacoustic signals. The main analytical expressions are provided for one-dimensional algorithm and its multidimensional modification. The main steps of software implementation of a multidimensional algorithm are provided for MATLAB.

Hydroacoustic signal, vector signal, weighted overlap add algorithm, software implementation

УДК 004.067

М. В. Лапаев
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)

А. И. Водяхо, А. Б. Смирнов, Н. А. Жукова
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Система обработки текстовых медицинских данных

Рассматриваются вопросы обработки текстовых медицинских данных. В записях врачей используется специализированная терминология, значительное число синонимичных понятий, содержится шум. Предложен алгоритм обработки медицинских текстов по шаблонам, который учитывает особенности данных. Алгоритм реализован и апробирован в составе системы семантической обработки и анализа медицинских данных, разрабатываемой для ФГБУ «СЗФМИЦ им. В. А. Алмазова» Минздрава России.

Семантические технологии, структурный анализ текста, текстовые медицинские записи

Проблема обработки данных на естественном языке актуальна для многих предметных областей, в частности – медицины, где значительная часть информации содержится в записях медицинских работников. Особенность текстовых данных в медицине заключается в использовании специализированной профессиональной терминологии. К настоящему времени проведены исследования и разработаны методы для автоматизированного распознавания, машинного перевода и анализа клинических записей. Трудности обработки текстов обусловлены отсутствием структуры в тексте, отсутствием эталонных фрагментов текстов, а также наличием синтаксического шума, синонимии и неоднозначностей. Существует множество подходов к решению задач обработки текстов на естественном языке. Наиболее широко применяются алгоритмы, построенные на основе статистического анализа [1], в частности частотного, построения графовых моделей [2], использования обучаемых языковых моделей [3]. Большинство из них позволяют решать только конкретные узкоспециализированные задачи. На сегодняшний момент не существует готового инструмента для обработки тестовых медицинских данных на русском языке. Задача может быть решена за счет совместного использования нескольких методов и алгоритмов. В настоящей статье задача обработки и анализа текстов рассматривается в контексте реализации процессов врачебных и управленческих задач, решаемых в Федеральном северо-западном медицинском центре им. В. А. Алмазова (Центр). Разрабатывается набор инструментов и средств, решающий задачу

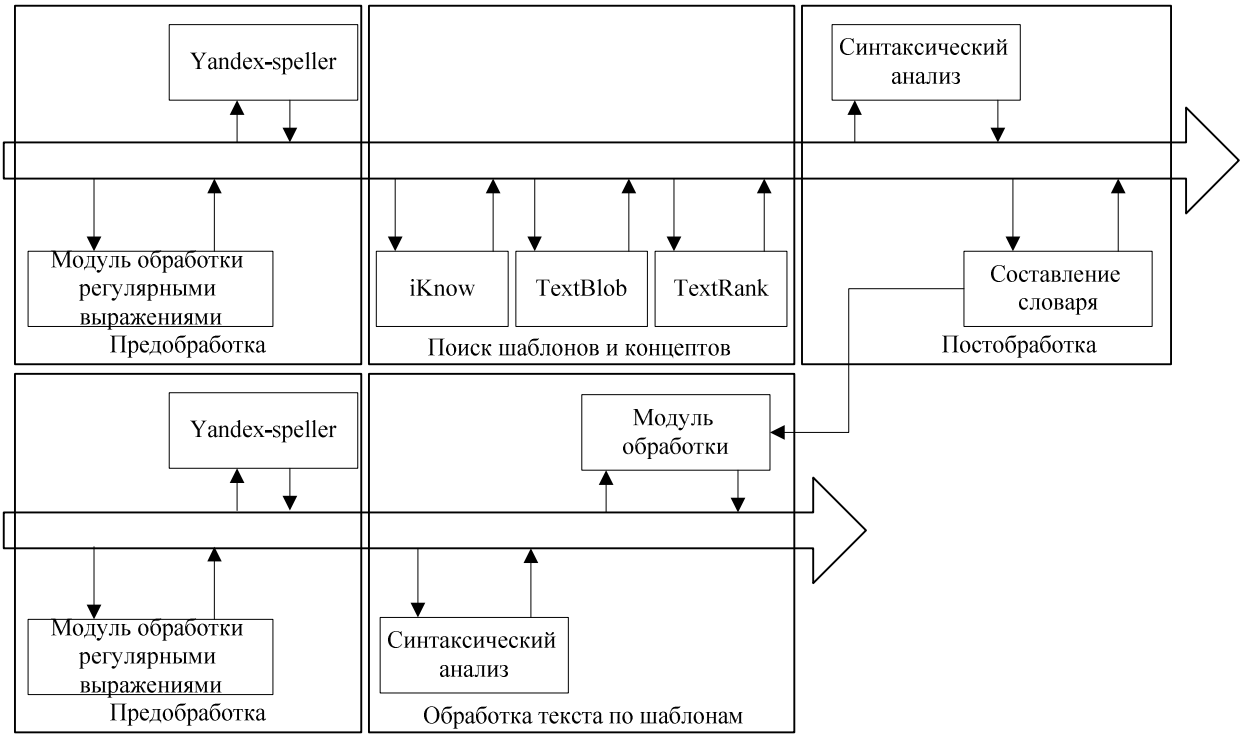

Рис. 1

предварительной очистки текста с учетом особенностей предметной области, а также задачу выделения концептов из записей с целью формирования пространства связанных данных о состоянии пациентов на основании шаблонной обработки. В состав медицинских данных входят численные данные (измеряемые результаты анализов, показатели диагностических тестов, снабженные единицами измерения, которые являются отдельной задачей обработки и сопоставления ввиду наличия нескольких обозначений для одной и той же величины), органолептические показатели (цвет, запах, прозрачность и т. п.) и субъективные текстовые записи (субъективность определяется наличием человеческого фактора). К основным типам текстовых записей относятся: анамнез жизни пациента, дневники и жалобы, рекомендации, диагнозы. Они составляют примерно 60 % от всех записей, количество которых исчисляется несколькими десятками миллионов записей. Также они несут наибольшую смысловую нагрузку среди записей для анализа.

Анализ реальных данных позволил выделить следующие особенности медицинских текстовых записей: наличие большого количества случайного шума, аббревиатур и сокращений, наличие человеческого фактора (различный стиль и структура письма, использование различной терминологии, наличие орфографических ошибок), практически полное отсутствие глаголов в структуре текста.

При обработке текстов на естественном языке обычно выделяют 4 уровня представления текстов [4]: символьный; лексический; синтаксический и семантический.

На *символьном* уровне текст документа рассматривается как последовательность символов. Для символов разработаны классификаторы. На *лексическом* уровне последовательность символов разделяется на слова. Класс, задающий множество слов, называют лексическим типом. Под лексическим типом понимается исходная словоформа и производные словоформы, которые образуются от исходной. Набор словоформ носит название парадигмы, а базовая словоформа – лексемы. На *синтаксическом* уровне в тексте выделяются предложения, анализируются связи между словами в предложении. Связи имеют иерархическую структуру и, как правило, представимы в виде дерева. На *семантическом* уровне строятся смысловые конструкции на основе выявленных в тексте семантических структур. Предлагаемая схема обработки медицинских текстов, в основу которой положена рассмотренная 4-уровневая модель, приведена на рис. 1.

Для корректной лемматизации (приведения словоформы к словарной форме) и последующего анализа теста требуется выполнить проверку орфографии. Проверка реализуется за счет использования сервиса Yandex-speller. Текст проверяется в соответствии с правилами орфографии и лекси-

ки современного языка. Орфографический словарь сервиса содержит информацию о написании большинства наиболее употребляемых слов. Для выявления таких ошибок, как «не_работает», «_безболезненный» и т. д., необходимо использовать регулярные выражения. Один из примеров обработки регулярными выражениями – удаление в тексте подчеркиваний и дублирующихся кавычек. С использованием регулярных выражений также решается задача расстановки пропущенных пробелов, устранения избыточных знаков препинания.

Эвристика – это алгоритм решения задачи, не имеющий строгого обоснования, но, тем не менее, дающий приемлемое решение в большинстве практически значимых случаев [5]. Эвристики используются для исправления таких синтаксических ошибок, как пропущенная точка в конце предложения, использование строчной буквы в начале предложения и т. п. В алгоритме обработки и анализа медицинских текстов для поиска и последующего исправления синтаксических ошибок используются регулярные выражения. Задачи по исправлению синтаксических ошибок могут также решаться с использованием таких инструментов, как DroolsGuvnor.

Шаблон в контексте данной работы – это устойчивая, повторяющаяся языковая конструкция, имеющая определенный семантический смысл и синтаксическую структуру. Для поиска шаблонов применяются методы и инструменты семантического анализа. Предлагается использовать 3 различных подхода к выявлению шаблонов. Подходы основаны на использовании инструментов iKnow (InterSystems Corporation Cambridge, MA [http://www.intersystems.com/ru/]), TextBlob (Open-source, [http://textblob.readthedocs.io/en/dev/]) и реализации метода TextRank [6].

Блок-схема алгоритма показана на рис. 2.

Инструмент iKnow позволяет представлять входной текст в виде последовательности предложений, разделенных знаками пунктуации. Анализ предложений выполняется с использованием конкретной языковой модели. Результатом анализа является последовательность, содержащая основные понятия и отношения между ними, представленные в виде кортежей <Concept, Relation, Concept> (CRC, <Сущность, Связь, Сущность>), и набор индексов, определенных на основе статистических показателей [7]. Под связью понимается слово или группа слов, логически соединяющая 2 понятия. На основе множества сущностей, определяемых iKnow, формируются понятия предметной области. Для формирования множества понятий привлекались эксперты предметной области, выполнялось сопоставление сущностей, выявлялись синонимы. Шаблоны описываются в терминах понятий предметной области. В отличие от других инструментов iKnow поддерживает встроенные возможности семантического анализа, т. е. входными данными является необработанный текст, выходными – CRC-тройки и концепты, которые используются для выявления си-

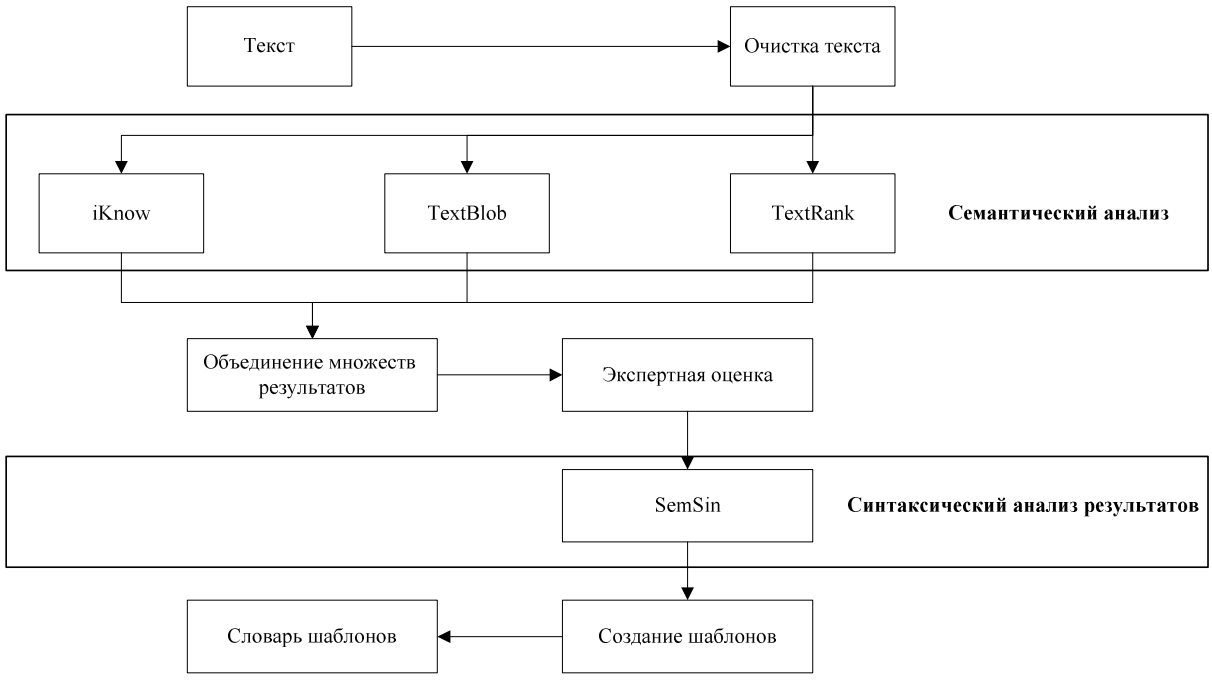

Рис. 2

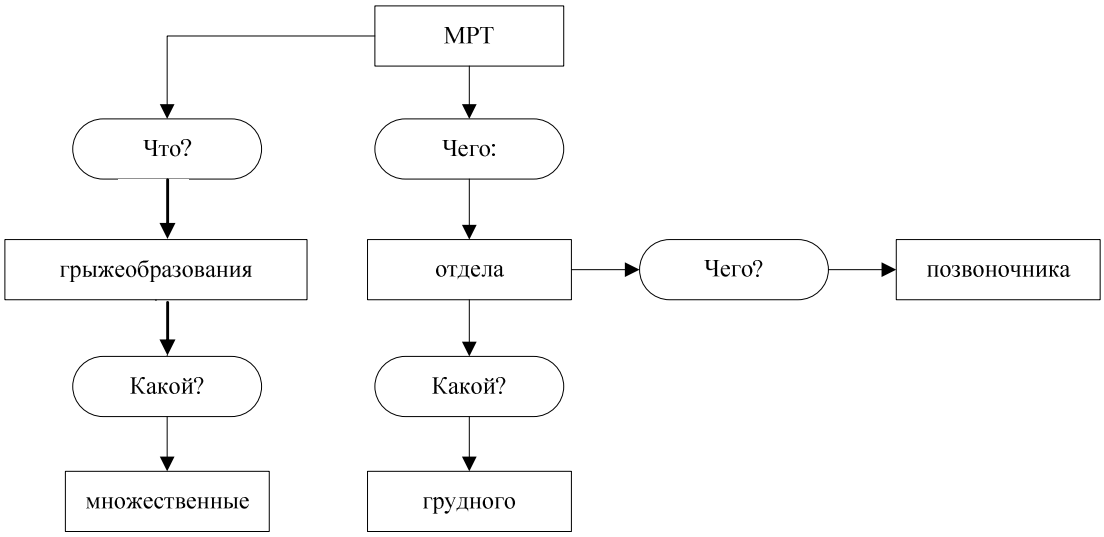

Рис. 3

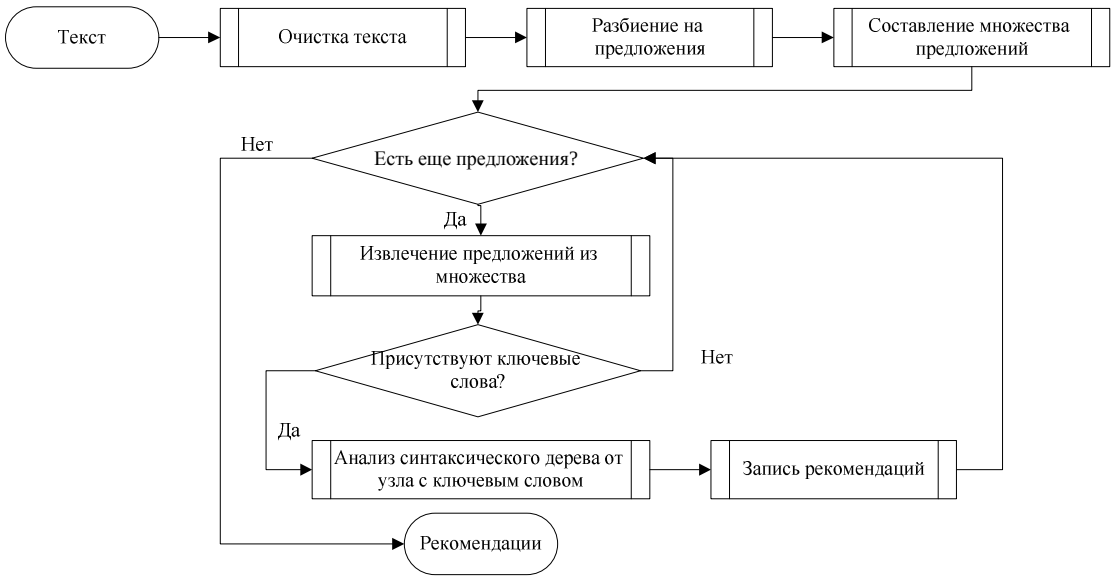

Рис. 4

нонимов среди них. Например, «за грудиной» и «в области сердца» будут синонимами, которые потом используются при создании шаблонов.

Инструмент TextRank основан на применении алгоритма PageRank, модифицированного для решения задач обработки естественного языка. Величина TextRank показывает значение стационарного распределения случайного блуждания для каждой вершины с учетом весов ребер [8]. Инструмент TextBlob представляет собой библиотеку, построенную на основе библиотек NLTK и Pattern (одна из разновидностей регулярных выражений). В библиотеке реализована функция поиска словосочетаний. Алгоритм работы с библиотекой аналогичен предложенному алгоритму для работы с iKnow. В результате применения алгоритма формируется множество концептов.

После нахождения словосочетаний эксперты выполняют проверку и, при необходимости, корректируют состав словосочетаний, исключают избыточные. Для оставшегося множества словосочетаний предусмотрен синтаксический анализ. Задача синтаксического анализа решалась с использованием готового инструмента – SemSin [9]. Анализатор SemSin позволяет строить синтаксическое дерево для каждого предложения и предоставляет результат обработки в виде xml-документа.

Пример отображения синтаксического дерева приведен на рис. 3.

После нахождения семантически значащих сочетаний и их синтаксического анализа объединяются с помощью найденных синонимичных концептов синтаксические конструкции, имею-

щие схожий семантический смысл. Полученные в результате конструкции являются шаблонами, из которых впоследствии формируется словарь. Блок-схема алгоритма представлена на рис. 4.

Для поиска шаблонов формируются правила. В левой части правил указываются ключевые слова. При выявлении ключевых слов проводится анализ синтаксического дерева для этого предложения, начиная с узла, соответствующего ключевому слову. В случае совпадения с шаблоном выдается рекомендация, связанная с ним.

Целью эксперимента является доказательство применимости алгоритма поиска шаблонов в текстовых данных, а также проверка его эффективности.

Для исследования анализировался тип записей «жалобы». Были взяты пациенты, имеющие в рекомендациях «суточное кардиомониторирование». Был использован инструмент Textblob для поиска словосочетаний. Среди них были выявлены схожие по смыслу.

Применение данной системы исследуется в центре «СЗФМИЦ им. В. А. Алмазова». Система организована таким образом, что в процессе своей работы она использует экспертные знания для расширения словаря шаблонов. Такая организация позволит экспертам корректировать словарь, а следовательно, учитывать различия стилистик записей и решать проблемы количества различных терминов.

В состав системы входят 2 компонента: компонент анализа частых последовательностей, предназначенный для поиска шаблонов и работы с ними, и компонент извлечения, который позволяет выполнить обработку текста по шаблонам и извлечь необходимые отношения из предложения. На вход системы поступают необработанные текстовые данные. Предобработка данных делится на 2 этапа. На первом этапе проводится предобработка входной последовательности с помощью регулярных выражений и эвристик, на втором выполняется проверка и исправление орфографических ошибок с помощью Yandex-speller. Разработанная система имеет производительность, достаточную для потоковой обработки врачебных записей в рамках проекта, что позволяет включить ее в цепочку процессов обработки для сквозной (от сырых данных к набору триплов) обработки медицинских записей на русском языке.

Результаты апробации системы на данных Центра им. Алмазова показали, что предложенный набор методов является достаточно точным для решения задач обработки медицинских записей. В настоящее время специалистами Центра ведется активная работа по наполнению словаря шаблонов, а также по дальнейшему тестированию метода.

СПИСОК ЛИТЕРАТУРЫ

1. Miner G. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

2. Cambria E., White B. Jumping NLP curves: a review of natural language processing research [review article] // Computational Intelligence Magazine. IEEE. 2014. Т. 9, № 2. P. 48–57.

3. Sidorov G. Syntactic n-grams as machine learning features for natural language processing //Expert Systems with Applications. 2014. Т. 41, № 3. P. 853–860.

4. Ахманова О. С. Словарь лингвистических терминов. М.: Едиториал УРСС, 2004.

5. Гудман С., Хидетниеми С. Введение в разработку и анализ алгоритмов. М.: Мир, 1981.

6. The PageRank Citation Ranking: Bringing Order to the Web / L. Page, S. Brin, R. Motwani, T. Winograd. 1999. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768.

7. Пат. США № 8417711. Data analysis based on data linking elements / M. Rik F. Brands, Van Hyfte, D. H. Medard. URL: http://freepatentsonline.com/8417711.html.

8. Усталов Д. Извлечение терминов из русскоязычных текстов при помощи графовых моделей. Екатеринбург: Изд-во УрФУ, 2012.

9. Боярский К. К., Каневский Е. А. Семантико-синтаксический парсер SemSin // Науч.-техн. вестн. информационных технологий, механики и оптики. 2015. Т. 15, № 5. С. 869–876.

M. V. Lapaev
Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University)

A. I. Vodyaho, A. B. Smirnov, N. A. Zhukova
Saint Petersburg Electrotechnical University «LETI»

INFORMATION SYSTEM FOR MEDICAL RECORDS PROCESSING

The paper is focused on the problem of medical records processing. Doctors' records include specific terminology, a significant number of synonymous terms and random text noise. We propose an algorithm of medical records processing using patterns. The algorithm takes into account the specific features of data. The algorithm is implemented and tested as a module of SMDA (Semantic medical data analysis) system for Almazov medical research center in St. Petersburg. Evaluation of the results was carried out by experts of the domain.

Semantic technologies, structural text analysis, medical records

УДК 681.5.015, 519.876.2

А. В. Экало, С. А. Кудряков, Е. Н. Шаповалов, Ю. Б. Остапченко, С. А. Беляев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Алгоритм принятия обоснованных решений в нештатных ситуациях на основе моделей нечетких множеств

Описана модель развития нештатной ситуации, анализ и обоснование критериев принятия решения по выходу из нештатной ситуации с использованием нечетких множеств. Разработан обобщенный алгоритм принятия решения, приведен пример его применения.

Нерасчетная нештатная ситуация, антагонистическая игра, математическая модель, нечеткая логика, алгоритм принятия решений, авиационная и ракетно-космическая техника

Современные задачи повышения эффективности и обеспечения безопасности эксплуатации сложных технических систем предъявляют повышенные требования к уровню профессиональной подготовки специалистов. Однако реалии современной действительности показывают снижение общего уровня подготовки выпускников средних школ и существенные изменения в социальной структуре мотивации к освоению будущей профессии. Это затрудняет процесс подготовки специалистов в профильных высших учебных заведениях, а недостаток материального обеспечения учебного процесса приводит к необходимости существенно увеличивать период адаптации молодых специалистов непосредственно в эксплуатирующих организациях.

В профессиональной подготовке специалистов по эксплуатации сложных технических систем (например, летного и диспетчерского состава авиации и специалистов ракетно-космического профиля) широко используются различные типы тренажерных систем.

Совершенствование профессионального мастерства должно сопровождаться тренировкой оперативного мышления специалиста. В связи с этим современные обучающие комплексы в качестве важнейшего условия должны иметь возможность имитировать проблемные и конфликтные ситуации.

В процессе эксплуатации комплексов авиационной и ракетно-космической техники (АРКТ) возникает достаточно большое количество неисправностей и других отклонений от установленных требований к протеканию и результатам технологических операций, называемых нештатными ситуациями (НС). Дальнейшее развитие НС может привести к происшествиям (катастрофам,