

# ИНФОРМАЦИОННО- УПРАВЛЯЮЩИЕ СИСТЕМЫ

НАУЧНЫЙ ЖУРНАЛ



4(47)/2010



4(47)/2010

РЕЦЕНЗИРУЕМОЕ ИЗДАНИЕ

# ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

**Учредитель**  
ОАО «Издательство «Политехника»»

**Главный редактор**  
М. Б. Сергеев,  
доктор технических наук, профессор

**Зам. главного редактора**  
Г. Ф. Мощенко

**Редакционный совет:**  
**Председатель** А. А. Оводенко,  
доктор технических наук, профессор  
В. Н. Васильев,  
доктор технических наук, профессор  
В. Н. Козлов,  
доктор технических наук, профессор  
Ю. Ф. Подоплекин,  
доктор технических наук, профессор  
Д. В. Пузанков,  
доктор технических наук, профессор  
В. В. Симаков,  
доктор технических наук, профессор  
А. Л. Фрадков,  
доктор технических наук, профессор  
Л. И. Чубраева,  
доктор технических наук, профессор, чл.-корр. РАН  
Р. М. Юсупов,  
доктор технических наук, профессор, чл.-корр. РАН

**Редакционная коллегия:**  
В. Г. Анисимов,  
доктор технических наук, профессор  
Е. А. Крук,  
доктор технических наук, профессор  
В. Ф. Мелехин,  
доктор технических наук, профессор  
А. В. Смирнов,  
доктор технических наук, профессор  
В. И. Хименко,  
доктор технических наук, профессор  
А. А. Шалыто,  
доктор технических наук, профессор  
А. П. Шепета,  
доктор технических наук, профессор  
З. М. Юлдашев,  
доктор технических наук, профессор

**Редактор:** А. Г. Ларионова  
**Корректор:** Т. В. Звертановская  
**Дизайн:** А. Н. Колешко, М. Л. Черненко  
**Компьютерная верстка:** С. В. Барашкова  
**Ответственный секретарь:** О. В. Муравцова

**Адрес редакции:** 190000, Санкт-Петербург,  
Б. Морская ул., д. 67, ГУАП, РИЦ  
Тел.: (812) 494-70-44  
Факс: (812) 494-70-18  
E-mail: 80x@mail.ru  
Сайт: www.i-us.ru

Журнал зарегистрирован в Министерстве РФ по делам печати, телерадиовещания и средств массовых коммуникаций.  
Свидетельство о регистрации ПИ № 77-12412 от 19 апреля 2002 г.

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук».

Журнал распространяется по подписке. Подписку можно оформить через редакцию, а также в любом отделении связи по каталогам: «Роспечать»: № 48060, № 15385; «Пресса России»: № 42476.

© Коллектив авторов, 2010

## ОБРАБОТКА ИНФОРМАЦИИ И УПРАВЛЕНИЕ

- Кипяткова И. С., Карпов А. А.** Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи 2  
**Воробьев С. Н., Гирина Н. В., Лазарев И. В.** Оценивание временного положения импульсного сигнала 9

## ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

- Костоготов А. А., Костоготов А. И., Чеботарев А. В.** Метод объединения принципа максимума в параметрических задачах оптимального управления 15

## МОДЕЛИРОВАНИЕ СИСТЕМ И ПРОЦЕССОВ

- Койгеров А. С., Дмитриев В. Ф.** Радиомаркер на поверхностных акустических волнах с помехоустойчивым частотно-манипулированным кодом 22  
**Селиванова Е. Н., Городецкий А. Е.** Компьютерное моделирование процессов возбуждения и синхронизации колебаний ресничек мерцательных клеток 29

## ПРОГРАММНЫЕ И АППАРАТНЫЕ СРЕДСТВА

- Солнцев Р. И., Тревогода М. А.** Программное обеспечение подсистемы САПР замкнутой системы управления «Природа-техногенника» 34  
**Суясов Д. И.** Выделение структурных признаков изображений символов на основе клеточных автоматов с метками 39  
**Михеева В. Д.** Методы расширения языков программирования (Часть 1) 46

## ЗАЩИТА ИНФОРМАЦИИ

- Григорьян А. К., Литвинов М. Ю.** Применение вейвлет-преобразования для внедрения ЦВЗ в видеопоток в режиме реального времени 53

## УПРАВЛЕНИЕ В МЕДИЦИНЕ И БИОЛОГИИ

- Кузнецов А. А.** Количество информации и энтропия ярусной диаграммы ритма сердца 57

## УПРАВЛЕНИЕ В СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ

- Караев Р. А., Сафарли И. И., Нагиев М. А., Абдурагимов Т. Ф., Гюльмамедов Р. Г.** Когнитивный анализ и управление инновационными проектами предприятий 63  
**Тушавин В. А.** Менеджмент качества службы поддержки пользователей в области информационных технологий 69  
**Карасев В. В., Соложенцев Е. Д.** Тематика исследований по логико-вероятностному управлению риском и эффективностью в структурно-сложных системах 72

## КРАТКИЕ СООБЩЕНИЯ

- Курбанов В. Г.** Метод оценки надежности сложных технических систем 75

## СВЕДЕНИЯ ОБ АВТОРАХ

77

## АННОТАЦИИ

82

ЛР № 010292 от 18.08.98.  
Сдано в набор 20.05.10. Подписано в печать 11.08.10. Формат 60х84/8.  
Бумага офсетная. Гарнитура SchoolBookC. Печать офсетная.  
Усл. печ. л. 11,0. Уч.-изд. л. 14,0. Тираж 1000 экз. Заказ 267.

Оригинал-макет изготовлен в редакционно-издательском центре ГУАП.  
190000, Санкт-Петербург, Б. Морская ул., 67.

Отпечатано с готовых диапозитивов в редакционно-издательском центре ГУАП.  
190000, Санкт-Петербург, Б. Морская ул., 67.

УДК 004.522

# АВТОМАТИЧЕСКАЯ ОБРАБОТКА И СТАТИСТИЧЕСКИЙ АНАЛИЗ НОВОСТНОГО ТЕКСТОВОГО КОРПУСА ДЛЯ МОДЕЛИ ЯЗЫКА СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

И. С. Кипяткова,

младший научный сотрудник

А. А. Карпов,

канд. техн. наук, старший научный сотрудник

Санкт-Петербургский институт информатики и автоматизации РАН

Описывается процесс автоматической обработки текстового корпуса, собранного из новостных лент ряда интернет-сайтов, для создания вероятностной  $n$ -граммной модели разговорного русского языка. Приводится статистический анализ данного корпуса, даются результаты по подсчету частоты появления различных  $n$ -грамм слов. Представлен обзор существующих типов статистических моделей языка.

**Ключевые слова** — модель языка, текстовый корпус русского языка, автоматическая обработка текста.

## Введение

Для генерации грамматически правильных и осмысленных гипотез произнесенной фразы распознавателю речи необходима некоторая модель языка или грамматика, описывающая допустимые фразы. Процесс распознавания речи может быть представлен как поиск наиболее вероятной последовательности слов [1]:

$$W = \arg \max_W P(W | A) = \arg \max_W P(A | W)P(W),$$

где  $P(A|W)$ ,  $P(W)$  — вероятности появления гипотезы по оценке акустической и языковой модели соответственно.

Для многих языков (например, английского) разработаны методы создания моделей языка, которые позволяют повысить точность распознавания речи. Но эти методы не могут быть напрямую применены для русского языка из-за свободного порядка слов в предложениях и наличия большого количества словоформ для каждого слова.

Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе  $n$ -грамм слов, цель которой состоит в оценке вероятности появления цепочки слов  $W = (w_1, w_2, \dots, w_m)$  в некотором тексте.

$n$ -граммы представляют собой последовательность из  $n$  элементов (например, слов), а  $n$ -граммная модель языка используется для предсказания элемента в последовательности, содержащей  $n - 1$  предшественников. Эта модель основана на предположении, что вероятность какой-то определенной  $n$ -граммы, содержащейся в неизвестном тексте, можно оценить, зная, как часто она встречается в некотором обучающем тексте.

Вероятность  $P(w_1, w_2, \dots, w_m)$  можно представить в виде произведения условных вероятностей входящих в нее  $n$ -грамм [2]:

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, w_2, \dots, w_{i-1})$$

или аппроксимируя  $P(W)$  при ограниченном контексте длиной  $n - 1$ :

$$P(w_1, w_2, \dots, w_m) \cong \prod_{i=1}^m P(w_i | w_{i-n+1}, w_{i-n+2}, \dots, w_{i-1}).$$

Вероятность появления  $n$ -граммы вычисляется на практике следующим образом:

$$P(w_i | w_{i-n+1}, \dots, w_{i-1}) = \frac{C(w_{i-n+1}, \dots, w_i)}{C(w_{i-n+1}, \dots, w_{i-1})},$$

где  $C$  — количество появлений последовательности в обучающем корпусе.