Секция
«Информационные технологии в образовании и производстве»

А.Д. Варламов
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: varlamov_aleks@mail.ru

Проблемы практического применения наивного байесовского классификатора текстов

Одним из наиболее известных методов классификации текстов на основе машинного обучения является наивный байесовский классификатор (Naive Bayes classifier). Исследование публикаций показало, что подавляющее большинство современных работ [1-4], автоматически относящие текстовые документы к одной из заданных групп, в анализе подходов к реализации и сравнении результатов не обходят стороной традиционный байесовский метод. Кроме того, его нередко используют в основе своих алгоритмов. Следует отметить, что, хотя в практических задачах классификации текстов байесовская модель в чистом виде практически не используется, ее стараются комбинировать с другими методами (семантический анализ, деревья решений и другие) для улучшения характеристик алгоритма. Поэтому исследование проблем, связанных с построением алгоритмов классификации текстовых документов на основе байесовского подхода является актуальным.

Классификатор Байеса представляет собой классическую, теоретически обусловленную вероятностную модель, реализовать которую не составляет большого труда. Но, как известно, теоретически обусловленные модели не всегда пригодны для решения практических задач или их реализация сопровождается большими трудностями. В докладе описаны и обосновываются проблемы применения байесовского подхода к классификации текстов.

Проблема 1 связана с погрешностями статистических оценок и вычислений. В докладе показано, что обеспечение требуемого уровня точности оценки выполняется за время, пропорциональное третьей степени среднего количества слов в анализируемых текстах. С точки зрения теории алгоритмов вычислительная задача, имеющая полиномиальный порядок временной сложности третьей степени, является неэффективной и не желательной в реализации, если используются данные большого размера. Поэтому не рекомендуется использовать наивный байесовский классификатор текстов, если документы содержат большое количество слов.

Проблема 2. Учет редких слов в модели сильно искажает оценку принадлежности документа к группе. Если учитывать очень редкие слова, а также слова, в которых допущена ошибка или опечатка, могут быть получены совсем неадекватные результаты. Например, если в обучающей выборке некоторое слово встретилось только в текстах первой группы, а другое только во второй (это вполне возможно для слов с очень низкой частотой встречаемости), и эти два слова встречаются в новом тексте, который следует отнести либо к первой группе, либо ко второй, то при байесовской классификации возникает неопределенность типа 0/0.

Проблема 3. Не удобная программная реализация. Среди стандартных агрегирующих функций в SQL-запросах к базам данных нет произведения (как в прочем и во многих инструментах работы со структурированными данными). Поэтому единым запросом к базе не может быть получено произведение выбранных значений ни в исследовательских проектах, ни в рабочих вариантах системы. Это ведет к усложнению логики программного кода и времени работы алгоритмов.

Чтобы избежать указанных проблем предлагается разбивать тексты на фрагменты (например, на абзацы, на предложения или сочетания слов) и найти вероятности принадлежности каждого фрагмента к группам. Затем определить общие статистические оценки принадлежности текстового документа к классам.

В проведенном исследовании реализованных алгоритма Байеса и алгоритма, усредняющего байесовские оценки для предложений, показаны преимущества второго. Значения вычисленных метрик полноты, точности, аккуратности и F-меры для второго алгоритма оказались лучшими.

Литература

1. Васильев В.Г., Худякова М.В., Давыдов С. Классификация отзывов пользователей с использованием фрагментных правил // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11: В 2 т. Т. 2 – М: Изд-во РГГУ, 2012. С. – 66-76.
2. Зайцев В.Г., Лан Чуньлинь. Способы повышения эффективности классификации документов для конечного множества языков // Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка. №50. 2010.
3. Калинин А.В. Применимость Байесовского классификатора для задачи определения спама // Материалы конференции "Проблема спама и ее решения". Москва, 2004.
4. Poroshin V. Proof of concept statistical sentiment classification at ROMIP 2011 // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18). – М.: Изд-во РГГУ, 2012. – С. 60–65.

С.В. Еремеев
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23

Алгоритмы учета муниципальных объектов для трехмерной ГИС города

Геоинформационные системы (ГИС) в последнее время находят все большее применение. Особенно интенсивно развиваются муниципальные ГИС, где наблюдается создание подробных карт городских объектов, применение GPS для контроля за движущимися объектами, обработка видеоизображений с автомобильных трасс и т.д. То есть происходит расширение функционального состава городской ГИС.

Создание трехмерной ГИС является естественным продолжением в развитии системы управления городом. На данный момент уже существуют трехмерные модели некоторых городов, к которым можно отнести, например, Москву, Казань, Санкт-Петербург. Эти модели могут быть созданы с помощью различных систем трехмерного моделирования: AutoCad, 3ds max, ArcGIS 3D Analyst, google sketchup. Результатом работы является векторная трехмерная модель города.

Для центральной части города Муром разработана трехмерная модель с помощью таких программных приложений как google sketchup и ArcGIS 3D Analyst. При этом рассматривается трехмерное построение следующих типов объектов: муниципальные объекты недвижимости, жилые строения, дорожные покрытия, кварталы, наземные инженерные коммуникации.

Между такими объектами также как и для двумерного случая распространяются топологические отношения, но в более сложном формальном описании с геометрической точки зрения. Наибольший интерес представляют отношения между строениями и коммуникациями. Эти отношения классифицированы и формально описаны.

Фрагмент 3D модели города Муром показан на рисунке 1.

Рис. 1. Фрагмент карты с размещенными точечными объектами, расположенными на остановках

В трехмерной ГИС города предусмотрено выполнение следующих задач:

1. Планирование строительства новых объектов на заданной территории и сравнение текущих результатов строительства с запланированными трехмерными моделями.

2. Автоматизация документооборота имущественных объектов с подвязкой к трехмерной модели объекта.

3. Анализ состояния дорожного покрытия территории города.

4. Учет, ремонт и анализ инженерных коммуникаций города.

5. Моделирование дорожных ситуаций с учетом трехмерного расположения различных объектов.

Данные алгоритмы расширяют возможности муниципальной ГИС, а также позволяют по запросу оперативно и наглядно получить необходимый результат.

Работа выполнена при финансовой поддержке РФФИ (проект № 12-07-31182 мол_а)

Е.Е.Канунова
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: kanunovaee@list.ru

Структура системы управления информационными ресурсами регионального музея

Исследования в области разработки алгоритмов управления ресурсами музеев, методов и систем автоматизированной реставрации цифровых изображений архивных документов являются весьма актуальными. Это связано с развитием новых информационных технологий и, как следствие, появлением новых подходов к хранению и использованию исторических и архивных текстовых документов.

В настоящее время большинство музеев, как в России, так и за рубежом участвуют в программах оцифровки и копирования уникальных фондов. Процессы внедрения в музеи электронных коллекций приводят к необходимости создания баз видеоданных, систем удобного хранения и распределения видеоданных, а также автоматизации реставрации графических данных, в частности изображений текстовых и фотографических документов.

Автором была разработана система управления информационными ресурсами музея, структурно-функциональная организация (СФО) которой описана в [1].

Основными составляющими СФО системы управления являются:

1. Модуль блоков формирования, оценки и учета информационных ресурсов регионального музея [2];
2. Модуль блоков управления информационными ресурсами;
3. Модуль блоков автоматизированной реставрации изображений архивных текстовых документов (АТД), а также распознавания старопечатных и скорописных символов [3,4].

В докладе подробно рассмотрены особенности разработки и исследований алгоритмов управления информационными ресурсами музея, методов реставрации, являющихся основой для создания систем хранения, распределения и реставрации документов.

Особое внимание уделяется модулю блоков управления информационными ресурсами музея. Он состоит из блока формирования критериев оценки музейного предмета, блока управления размещением изображений, блока обработки запросов пользователей и блока получения/передачи изображений.

Последние три блока образуют подсистему хранения и распределения изображений музейных материалов (видеоданных), отличительной особенностью которой является трехкомпонентная структура организации и хранения видеоданных, позволяющая увеличить скорость доступа к электронным коллекциям музея. Для реализации процесса управления информационными ресурсами регионального музея, а так же для обеспечения сопряжения подсистем системы управления информационными ресурсами разработан блок – менеджер системы.

Литература

1. Макарова Е.Е. (Канунова Е.Е.), Варламов А.Д. Структурно-функциональная организация системы управления информационными ресурсами регионального музея // Алгоритмы, методы и системы обработки данных. 2011. №2. – С. 50-55.
2. Садыков С.С., Канунова Е.Е. Система формирования данных об информационных ресурсах краеведческого музея и управления ими: опыт разработки и использования // Информационные технологии. 2007. №10. – С.59-65.
3. Садыков С.С., Канунова Е.Е., Варламов А.Д. Автоматизированная реставрация изображений архивных текстовых и фотографических документов // Автоматизация и современные технологии. 2007. №8. – С.10-12.
4. Канунова Е.Е., Полякова Е.В. Особенности распознавания изображений старопечатных текстовых символов // Алгоритмы, методы и системы обработки данных. 2009. №14. – С. 55-61.

А.Н.Коноплев
Муромский институт (филиал) Владимирского государственного университета
602200, г. Муром Владимирской обл., ул. Орловская, 23,
E-mail: kaf-eivt@mivlgu.ru

Разработка программного комплекса автоматизированной системы оперативно-календарного планирования

В данном докладе рассматривается задача разработки программного обеспечения для автоматизированной системы составления оперативно-календарного плана методами минимизации времени технологического цикла для мелкосерийного радиоэлектронного производства.

Качество и эффективность функционирования производственного процесса предприятия во многом определяется эффективностью оперативно-календарного планирования. Задачи оперативно-календарного планирования отражают процесс распределения во времени ограниченного числа ресурсов для выполнения проекта, состоящего из заданного множества взаимосвязанных работ. Процесс организации оперативно-календарного планирования достаточно трудоемкий и требующий существенных временных затрат, в то же время существующие технологии составления оперативно-календарных планов не всегда в состоянии отслеживать динамику изменения условий организации производства. В этих условиях наиболее актуальной задачей является разработка автоматизированной системы построения оперативно-календарных планов. Задачей данного исследования является разработка информационного обеспечения и структурной схемы программного обеспечения для автоматизированной системы построения оперативно-календарного плана мелкосерийного производства радиоэлектронных изделий с учетом всех технологических условий производства.

Информационное обеспечения должно быть построено в соответствии с требованиями к функциональности создаваемой автоматизированной системы [1]. Основные функции, которые должна выполнять автоматизированная система оперативно-календарного планирования для изготовления радиоэлектронных изделий это:

– хранение, поиск, выдача информации, необходимой в процессе планирования;

– создание контрольных отчетов, содержащих выполнимый план-график (сменно-суточные задания) работы технологического оборудования;

– построение и перестроение план-графика по критериям оптимальности заданным диспетчером;

– вывод справочной информация (о технологических процессах) или текущей информации (состояние производства на данный момент) на печать в виде таблиц или диаграммы Ганта;

– разграничение прав доступа пользователей в процессе использования системы;

– проверка на достоверность и выполнимость разработанных план-графиков (сменно — суточных планов) в регламентированные сроки.

Для реализации представленных функций в разрабатываемой автоматизированной системе предлагается ввести модульную структуру, структурная схема программного обеспечения представлена на рисунке 1, и состоит:

1.База данных с нормативно-справочной информацией. Данный модуль представляет собой сетевую базу данных, в которой хранится нормативно-справочную информация необходимая для планирования. Также данный модуль обеспечивает взаимосвязь с клиент-серверным приложением «Plan».

2.Модуль «Plan». Данный модуль является связующим модулем автоматизированной системы. Модуль является интерфейсом связи с пользователем, а также связи с остальными частями системы. В данном модуле задаются основные параметры системы: критерии оптимальности, параметры необходимые для построения расписания, критерии работы алгоритмов, и т.д.

3.Модуль «Work Plan». Модуль «Work Plan» является модулем расчета производственного календарного плана, на основе разработанного алгоритма из [2]. Данный модуль может выполнять следующие функции и процедуры:

– процедура назначения приоритетов и формирования массивов технологических операции (матрицы обработки деталей);

– процедура выбора и закрепления технологических операций за оборудованием;

– процедура расчета общего времени «пролёживания деталей» и общей длительности производственного план-графика;

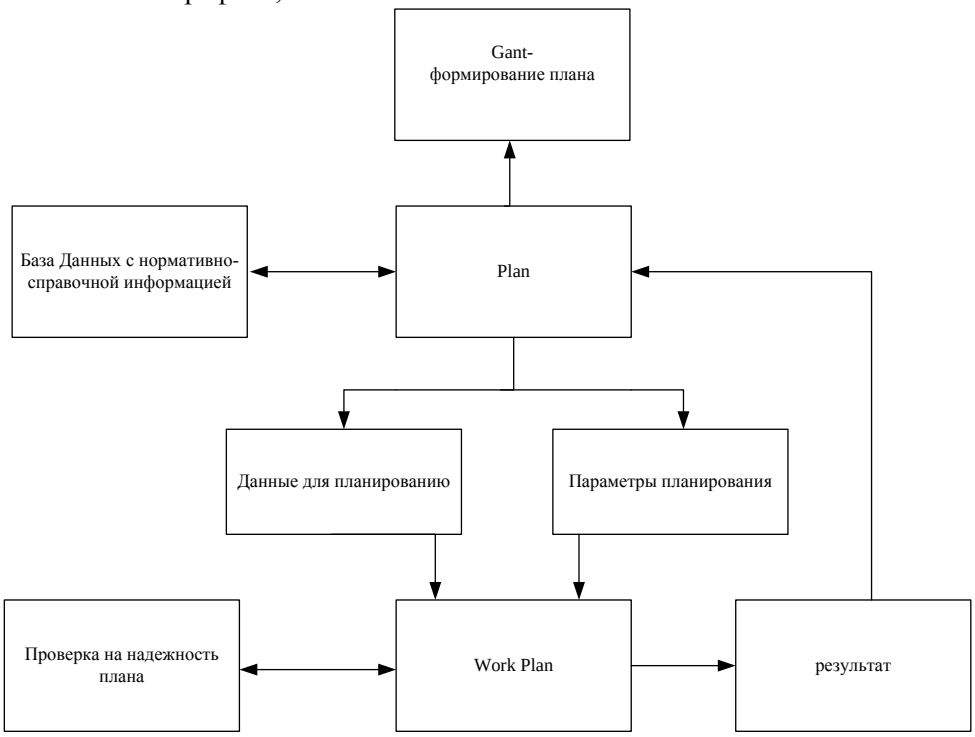

Рис.1. Структура программного обеспечения разработанной автоматизированной системы оперативно-календарного планирования

– процедура оценки построенного расписания в соответствие с выбранными критерием оптимальности

– процедура расчета дат и времени, осуществляет расчет добавления операций с проверкой графика допустимости работ и т.п.

4. Модуль «Gant». Данный модуль представляет собой программное обеспечение по формированию план-график работы представленной в виде диаграммы Ганта.

5. Модуль «Проверка на надежность плана». Является модулем проверки, реализует методику проверку достоверности разработанного плана из [3]. Если сформированный план в модуле «Work Plan» невыполним в регламентированные сроки, то пользователю будет предложено выбрать другие критерии и параметры разрабатываемого плана.

6. Модуль «Результат». Данный модуль представляет собой полностью сформированный по заданным критериям и проверенный на выполнимость план-график работы технологического оборудования.

Таким образом, разрабатываемая система позволяет автоматизировать процесс разработки календарных план-графиков на мелкосерийном производстве радиоэлектронных изделий.

Литература

1. Коноплев, А.Н. Разработка функциональной модели для системы оперативно-производственного планирования /Коноплев А.Н.// Материалы XVIII Международной конференции по вычислительной механике и современным прикладным программным системам, 22-31 мая 2013 г., Алушта. – М.:Изд-во МАИ, 2013. – С.769-771

2. Коноплев, А.Н. Алгоритм оперативно-календарного планирования мелкосерийного производства / Коноплев А.Н., Кропотов Ю.А.// Автоматизация в промышленности, 2013. №11. – С48-51.
3. Коноплев, А.Н. Математическая модель диагностики и восстановления технологического оборудования в мелкосерийном производстве / А.Н. Коноплев, Г.П. Суворова // Информационные системы и технологии, 2013. №3 (77). – С.30-36.

И.Л. Лушпа, М.А.Монахов
Московский институт электроники и математики Национального исследовательского университета «Высшая школа экономики»
E-mail: Mixa.monahov@list.ru

Разработка базы данных для оценки безотказности радиоэлектронной аппаратуры с учетом механических и электромеханических элементов

Уровень качества вновь создаваемой и модифицируемой радиоэлектронной аппаратуры, который определяет ее конкурентоспособность на внешнем и внутреннем рынке, в значительной степени зависит от эффективности и качества её проектирования. Современная аппаратура характеризуется сложными алгоритмами функционирования, обладает повышенной надёжностью, высокими удельными показателями, помехозащищённостью и стойкостью к широкому спектру внешних воздействующих факторов.

На надежность радиоэлектронной аппаратуры (РЭА) влияют механические воздействия.[1] Для снижения уровней этих воздействий применяются системы амортизации, в состав которых входят разнообразные механические и электромеханические элементы (М/ЭМ). В расчете надежности РЭА надежность таких элементов учитывается с помощью моделей, приведенных в [2]. В плане оценки надежности М/ЭМ больший интерес представляют модели интенсивностей отказов, приведенные в американском стандарте NSWC-2011/LE10 [3], разработанного специалистами Кардерокской дивизии ВМФ США.

Анализ зависимостей, моделей интенсивностей отказов различных М/ЭМ [4-6] стандарта [3], позволил предложить следующую классификацию параметров и коэффициентов:

- Параметры ТУ;
- Параметры режима применения;
- Эмпирические коэффициенты;
- Физические константы.

Кроме того, поскольку в стандарте [3] используется англо-американская система единиц, то в классификацию необходимо дополнительно ввести коэффициенты пересчета в систему СИ.

Применяющиеся в настоящее время для расчётной оценки надёжности аппаратуры методики, основанные на использовании как отечественных программных средств (АСРН, ПК «АРБИТР», модуль «Надёжность» комплекса КОК и др.), так и зарубежных (модули «Reliability» CAD-систем и специализированные системы, такие как RAM Commander, Windchill, BlockSim и др.), позволяют лишь в отдельных случаях частично оценить показатели надёжности проектируемой аппаратуры. И при этом во всех этих программах отсутствуют базы данных по конструктивно-технологическим параметрам М/ЭМ, необходимым для расчётов характеристик надёжности, поэтому эти данные необходимо вводить «вручную».

В этом плане, пожалуй, единственным исключением является система АСОНИКА-К-СЧ программного комплекса АСОНИКА-К, в которой математические модели интенсивности отказов хранятся в базе данных, а интерфейс пользователя может быть модифицирован без изменения её программного кода [7, 8]. Поэтому при разработке концептуальной модели базы данных (БД) по параметрам М/ЭМ за основу была принята модель справочной части системы АСОНИКА-К-СЧ по параметрам ЭРИ.

Поскольку база данных параметров механических и электромеханических элементов была интегрирована базу данных системы АСОНИКА-К-СЧ, то для разработки модуля расчёта надёжности механических и электромеханических элементов применялись инструментальные средства, используемые для модификации системы АСОНИКА-К-СЧ и её базы данных [7, 8].

Модуль расчёта надёжности механических М/ЭМ предназначен для формирования исходной информации (ввода данных для расчёта и формирования SQL-запросов к БД). Программирование таких модулей в системе АСОНИКА-К-СЧ осуществляется с помощью специализированного языка, описание которого приведено в [7].

Концептуальная модель БД по параметрам М/ЭМ приведена на рис.1.

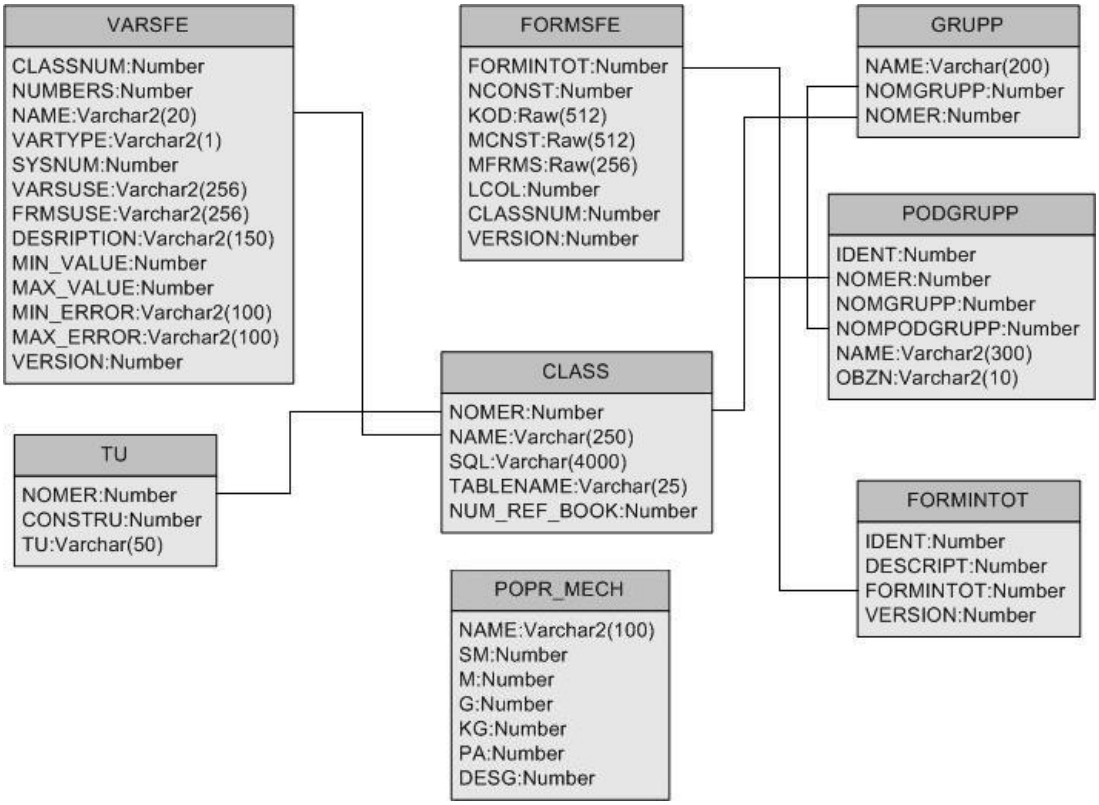

Рис. 1. Концептуальная модель базы данных

Как следует их рисунка 1, БД имеет иерархическую структуру (классы, группы, подгруппы М/ЭМ), соответствующую структуре стандарта [3].

Литература

1. Маркин, А.В. Методы оценки надёжности элементов механики и электромеханики электронных средств на ранних этапах проектирования. / А.В. Маркин, С.Н. Полесский, В.В. Жаднов. // Надёжность. 2010. № 2. – с. 63-70.
2. Справочник «Надежность электрорадизделий». – М.: МО РФ, 2006.
3. NSWC-2011/LE10. Handbook of reliability prediction procedures for mechanical equipment.
4. Лушпа, И.Л. Исследование модели интенсивности отказов изогнутых кольцевых пружин / И.Л. Лушпа, М.А. Монахов, В.М. Фокин. // Научные чтения по авиации посвященные памяти Н.Е.Жуковского. X Всероссийская научно-техническая конференция: сб. тез. докл. Всероссийской научно-технической конференции. Москва, 12 апр. 2013 г. – Москва: Издательский дом Академии имени Н.Е. Жуковского, 2013. [Электронный ресурс]: 1 электрон. опт. диск (CD-ROM).
5. Лушпа, И.Л. Исследование модели интенсивности отказов механических элементов класса «Пружины». / И.Л. Лушпа, М.А. Монахов, В.М. Фокин. // Инновационные информационные технологии: Материалы международной научно-практической конференции. Том 3. – М.:МИЭМ НИУ ВШЭ, 2013. – с. 443-446.
6. Лушпа, И.Л. Исследование модели интенсивности отказов волнообразных кольцевых пружин. / И.Л. Лушпа, М.А. Монахов, В.М. Фокин. // XXI Международная студенческая конференция-школа-семинар «Новые информационные технологии»: тез. докл. – МИЭМ НИУ ВШЭ, 2013.

7. Жаднов, В.В. Управление качеством при проектировании теплонагруженных радиоэлектронных средств: Учебное пособие. / В.В. Жаднов, А.В. Сарафанов. – М.: СОЛОН-ПРЕСС, 2012. – 464 с.

А.С. Платонова
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: allaplatonova@inbox.ru

Разработка пользовательского интерфейса информационной системы многопараметрического контроля образовательной деятельности

В рамках проектирования информационной системы многопараметрического контроля успешности освоения школьниками основных образовательных программ был разработан пользовательский интерфейс на основе web-шаблона Lightneasy с помощью PHP и JavaScript. Интерфейс – совокупность средств и правил, обеспечивающих взаимодействие устройств вычислительной системы и программ, а также взаимодействие их с человеком. С ростом сети Internet широкое распространение получили web-интерфейсы, позволяющие взаимодействовать с различными программами через браузер, который включает в себя основное меню, набор ссылок, реализующих те или иные функции системы и центральный блок, отображающий текущую задачу [1].

Набор возможностей разрабатываемого web-интерфейса зависит от пользователя, прошедшего аутентификацию. Для работы в информационной системе пользователь должен авторизоваться (ввести свой логин и пароль) или в случае нового пользователя заполнить форму регистрации. Возможности (набор функций) администратора и эксперта, учителя, ученика, родителей, административного персонала или гостя отличаются друг от друга. Непосредственно выполнению пользователями тех или иных задач им предшествует экран с инструкцией с данной функцией.

Интерфейс администратора и эксперта включает в себя следующие задачи:

– управление пользователями (активация, редактирование и удаление аккаунтов пользователя);

– функции редактирования учебного контента (добавление и редактирование тестовых заданий, методик и т.п.).

Интерфейс учителя включает в себя следующие задачи:

– контроль умения выполнять лабораторные и творческие работы, сформированность общеучебных умений и навыков и воспитанности учащихся (фактически это заполнение учителями электронных таблиц контроля);

– формирование оценок: генерация многопараметрической оценки достигнутого уровня результатов образования любого из учащихся или предоставление статистической информации в виде различного графиков или таблиц.

Интерфейс ученика предоставляет следующие возможности:

– выполнение тестовых заданий на проверку усвоения теоретического материала, решение контрольной работы, состоящей из нескольких задач, выполнение психологических тестов;

– формирование оценок, в частности генерирование многопараметрической оценки достигнутого уровня своих результатов образования.

Интерфейс родителей включает в себя задачу получения многопараметрической оценки достигнутого уровня результатов образования своего ребенка.

Интерфейс административного персонала (городские, районные и областные управления образования) позволит решать задачи предоставления многопараметрических оценок учащихся и статистической информации по ученику, классу, параллели, школе, городу и области.

Программная реализация разработанных процедурных моделей заключается в использовании в теле HTML страницы структур языка PHP и SQL-запросов. Несколько повторяющихся алгоритмических последовательностей (например, извлечение списка класса) выполнены в виде функций, которые можно вызывать неоднократно и с разным набором входных переменных. Кроме того, некоторые процедуры, в частности обработка результатов

тестирования, выполняется исполняемым PHP-скриптом в чистом виде без использования HTML.

Литература

1. Словарь терминов Интернет [Электронный ресурс]. – Режим доступа: http://your – hosting.ru/terms/h/hm/. Дата обращения: 19.10.2011.

С.С. Садыков, А.С. Белякова
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: asbelyakova@rambler.ru

Использование важности значения признаков при классификации заболеваний сердца

Патологии сердечно-сосудистой системы представляют собой угрозу жизни человека и являются причиной 57% летальных исходов. Верное определение заболевания является основным фактором, от которого зависит исход лечения[1]. Следовательно, разработка новых классификационных методик и модернизация уже существующих представляет собой востребованную сегодня задачу.

Любое сердечно-сосудистое заболевание (ССЗ) описывается множеством параметров, которые могут принимать одинаковые значения при различных патологиях. Поэтому задача их классификации является многопараметрической, а анализируемые признаки имеют различные типы (количественные, порядковые, качественные)[2,3]. Существующие методики определения ССЗ основываются на сравнении значений признаков пациентов с их эталонными интервалами при каждой из болезней. При этом считается, что если признак принадлежит интервалу, то это свидетельствует о наличии анализируемой патологии. Однако важность значения признака для данного заболевания не учитывается.

Алгоритм вычисления важности признака для данного заболевания состоит из следующих шагов:

1. Строится график значений эталонных интервалов признака x_1 при исследуемых ССЗ (рис. 1):

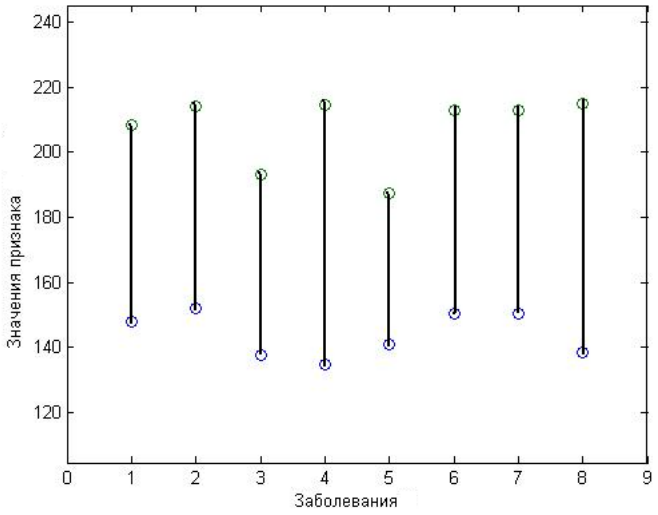

Рис. 1.

2. Строится график частоты встречаемости значений признака по:

$$G(x_{1_j}) = \sum_{i=1}^{m} \vartheta(x_{1_j}), \tag{1}$$

$$\vartheta(x_{1_j}) = \begin{cases} 1, \; x_{1_j} \in \left[x^{Ym}_{1\min} ; x^{Ym}_{1\max} \right], \\ 0, \text{èíà÷å} \end{cases} \tag{2}$$

где $x_{1_j} \in \left[x_{1\min} ; x_{1\max} \right]$, т.е x_{1_j} принимает все значения встречающиеся в обучающей выборке, Y_m - заболевание сердца, $m = \overline{1, M}$.

3. Важность значения признака пациента $x_{1,\text{їàä}}$ для определения каждого из исследуемых ССЗ рассчитывается по:

$$\varphi(x_{1\text{їàä}}^{Ym}) = \begin{cases} \dfrac{1}{G(x_{1,\text{їàä}})}, x_{1,\text{їàä}} \in \left[x_{1\min}^{Ym} ; x_{1\max}^{Ym} \right] \\ 0, \text{èíà÷å} \end{cases}. \quad (3)$$

Таким образом, при проведении классификации становится возможным учитывать важность признака для каждого заболевания конкретно, что позволяет сужать диапазон ССЗ среди которых производится отбор и сделать процесс принятия решения о наличии заболевания более объективным.

Разработка проводится совместно с врачами кардиологического отделения НУЗ Отделенческая больница на станции Муром ОАО РЖД.

В докладе рассматриваются вопросы применения предлагаемой методики оценки важности значений признаков при различных ССЗ на основе данных ПТК «КардиоВизор», индивидуальной информации о пациенте и данных инструментальных обследований.

Литература

1. Садыков С.С., Сафиулова И.А., Белякова А.С. Автоматическая объективная оценка и выбор наиболее значимых параметров для диагностики сердечно-сосудистых заболеваний // Автоматизация и современные технологии. 2012. №3. – с.27-33

2. Садыков С.С., Белякова А.С., Евстигнеева О.И., Жолобов С.А. Исследование взаимосвязи между окраской участков портрета сердца и изменениями электрокардиограммы // Приборостроение. 2012. №2. – с.64-69

3. Белякова А.С. Основные признаки оценки состояния сердчено-сосудистой системы // Алгоритмы, методы и системы обработки данных: сборник научных статей; Выпуск 14 / Под ред. С.С. Садыкова, Д.Е. Андрианова – М.: «Центр информационных технологий в природопользовании», 2009. – С.24-29.

М.В. Усачев, Н.П. Мольков
Муромский институт (филиал) Владимирского государственного университета
602264, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: um.mivlgu@gmail.com

Модели генерации случайных графов для социальных сетей

В настоящее время наблюдается значительный рост социальных сетей. Их размер варьируется от тысяч до миллионов пользователей. Наиболее известными из них являются сети Facebook, Twitter, Google+ и др. Прямые измерения динамики, отказоустойчивости и других характеристик социальных сетей затрудняются значительным временем и трудоемкостью сбора реальных данных. Одним из выходов является обработка снимков баз данных таких сетей, находящихся в свободном доступе. Однако, недостатком такого подхода является малое количество баз данных. Поэтому актуальным является математическое и имитационное моделирование социальных сетей в виде случайных графов.

Известно достаточное количество моделей генерации случайных графов. Их можно условно разделить на группы по признаку формирования графа [1] на структурно-управляемые, функционально-управляемые, намеренно-управляемые.

В работе [2] была рассмотрена адекватность графов Кронекера, модели Барабаши-Альберта, Random Walk, Nearest Neighbor, dK-графов, Forest Fire. Эксперименты проводились на базе данных Facebook от 2008 года, которая является репрезентативной и отражает основные характеристики, присущие также другим социальным сетям. Для оценки соответствия графа реальной социальной сети часто используются метрики: показатель степени закона распределения, коэффициент кластеризации, разделение вершин, клика и другие. Наиболее точными оказались графы Кронекера и dK-графы, однако их использование затруднено трудоемкостью процесса генерации. Другие модели обладают более высокой скоростью генерации графа, однако показывают меньшую точность в воспроизведении его характеристик.

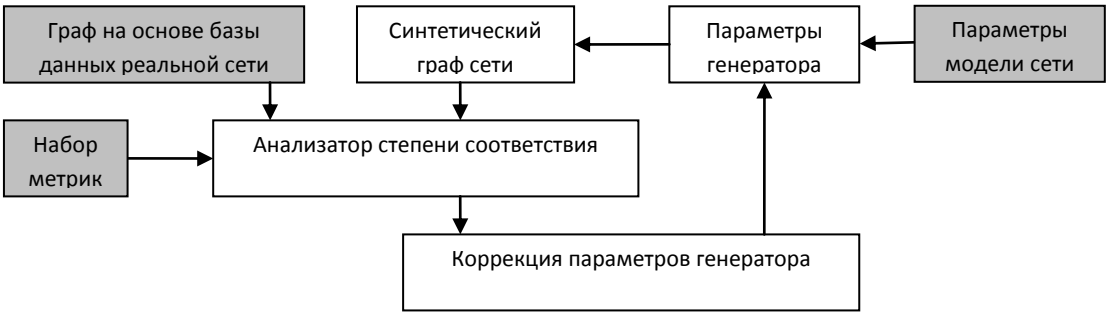

Рис. 1

В настоящей работе предлагается рассматривать подход к генерации графа как к задаче дискретной оптимизации его характеристик в соответствии с репрезентативной моделью реальной социальной сети. Возникает дополнительная задача выбора метрик, по которым будет определяться соответствие. Структурная схема процесса генерации синтетического графа представлена на рисунке 1. Серым цветом здесь отмечены исходные данные.

Литература

1. Shaozhi Ye, Juan Lang, Felix Wu. Crawling Online Social Graphs (англ.). — APWEB’12, April 6-8, 2010, Busan, Korea, 2010.
2. A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, B. Y. Zhao Measurement-calibrated Graph Models for Social Network Experiments (англ.). — WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA, 2010.

Е.С. Хлебнов
Муромский институт Владимирского государственного университета
602264 г. Муром, Владимирской обл., ул. Орловская, 23

Моделирование одноступенчатого алгоритма классификации с накоплением данных

Классическая формулировка задачи статистического синтеза радиоэлектронных систем заключается в максимизации или минимизации статистических критериев качества системы (вероятности ошибок, времени принятия решения и т.п.) при заданных ограничениях, налагаемых на саму систему или на воздействующие на нее сигналы и процессы. Применительно к распознающим системам наибольший интерес представляет следующий вариант постановки задачи оптимизации их характеристик: минимизируется суммарное количество наблюдений (определяемое объемом обучающих и контрольной выборок и размерностью признакового пространства), необходимое для обеспечения требуемого уровня достоверности распознавания при заданном наименьшем возможном расстоянии между классами [1]. С практической точки зрения представляет интерес задача оптимизации суммарного количества наблюдений (определяемое, в общем случае, объемом обучающих и контрольной выборок и размерностью признакового пространства), необходимого для обеспечения требуемого гарантированного уровня достоверности распознавания при заданном наименьшем возможном расстоянии между классами, в качестве которого из практических соображений естественно взять реальную точность измерения этого расстояния в распознающих системах.

Рассмотрим задачу определения принадлежности выборки, состоящей из n независимых нормально распределенных наблюдений, к одному из двух классов $A = \{a_1, a_2\}$. Измеряемые значения признака объекта x представляют собой реализации случайной величины с плотностью распределения

$$f(x, m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-m)^2}{2\sigma^2}}. \qquad (1)$$

Для построения решающего правила (алгоритма классификации объекта) по критерию максимального правдоподобия необходимо определить точки пересечения графиков условных плотностей вероятности $f(x|a_1) = f(x, m_1, \sigma_1)$ и $f(x|a_2) = f(x, m_2, \sigma_2)$ [2], т.е. значения признака x, для которых отношение правдоподобия равно единице:

$$L(x) = \frac{f(x, m_1, \sigma_1)}{f(x, m_2, \sigma_2)} = 1. \qquad (2)$$

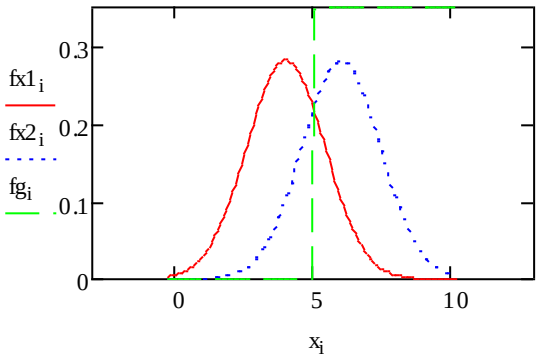

Рис. 1. Условные по классу плотности вероятности признака *x* и граница между классами

При совместной обработке совокупности значений измеряемого признака $\vec{x} = \{x_1, x_2, ...x_n\}$ в предположении независимости элементов x_i (i = 1, 2,…, n) выборки $\vec{x}$

условные по классу a_k функции правдоподобия (многомерные плотности распределения) будут иметь вид

$$f(x|a_k) = f(\vec{x}, m_k, \sigma_k) = \prod_{i=1}^{n} f(x_i, m_k, \sigma_k). \quad (3)$$

При априорно известных математических ожиданиях m_1, m_2 задача распознавания формулируется и решается в рамках теории проверки статистических гипотез как проверка простой гипотезы H_1: среднее значение нормально распределенных наблюдений равно m_1 против простой альтернативы H_2, что среднее равно m_2 при известной общей дисперсии σ. Решающая процедура заключается в сравнении логарифма отношения правдоподобия с некоторым порогом ln$c$, зависящим от выбранного критерия качества:

$$\ln L(\vec{x}) \geq \ln c, \quad a_2 > a_1 \quad (4)$$

При выполнении неравенства (4) принимается решение H_2, при невыполнении – H_1. Если решение принимается по критерию максимального правдоподобия (2), то ln$c$=ln1=0. Поскольку логарифм – монотонное преобразование [3], то достаточной статистикой для принятия решений будет среднее выборочное значений

$$y_n = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (5)$$

и для получения решающей распознающей процедуры (4) можно использовать правило

$$\gamma = \begin{cases} y_n > x_{gr} \Rightarrow H_2 : A = a_2 \\ y_n \leq x_{gr} \Rightarrow H_1 : A = a_1 \end{cases} \quad (6)$$

где порог принятия решения x_{gr} = ln$c$ определяется из (4) как

$$\ln c = x_{gr} = \frac{m_1 + m_2}{2}. \quad (7)$$

В работе проведено моделирование метода распознавания одномерных нормальных совокупностей с точки зрения оптимизации временных характеристик распознающей системы. Вычислен порог принятия решения и формализовано решающее правило.

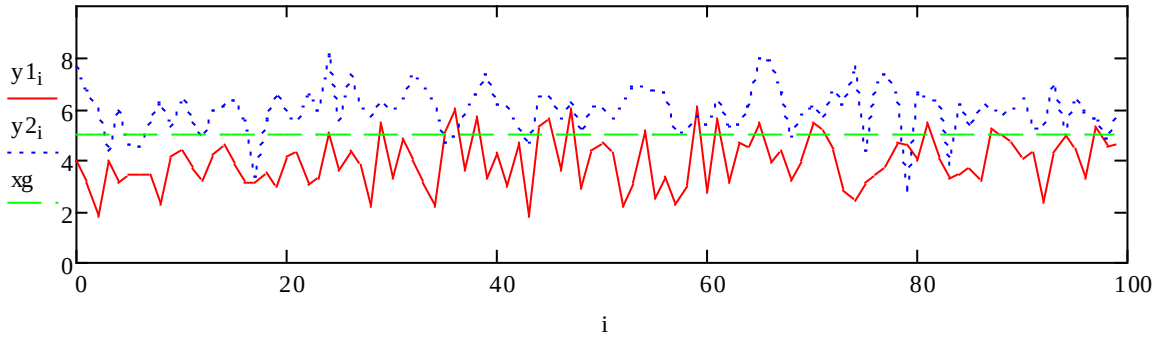

Рис. 2. Реализации статистик y_1 и y_2

Литература

1. Дуда Р. Распознавание образов и анализ сцен: пер. с англ. /Р. Дуда, П. Харт; под ред. В.Л. Стефанюка. – М.: Мир, 1976. – 511 с.
2. Фукунага, К. Введение в статистическую теорию распознавания образов. – М.: Наука, 1979.
3. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978.
4. Фомин Я.А. Статистическая теория распознавания образов /Я.А. Фомин, Г.Р. Тарловский. – М.: Радио и связь, 1986. – 264 с.

П.А. Цыганов
Московский институт электроники и математики Национального исследовательского университета «Высшая школа экономики»
109028, Москва, Б. Трехсвятительский пер., д. 3.
E-mail: tsyganov.p@gmail.com

Алгоритм работы пользователя с информационным порталом

При разработке современных информационных порталов [1-3] большое внимание необходимо уделять удобству работы с ними. Необходимо уделять внимание меню портала, оно должно быть простым. Любые разделы портала должны быть доступны по переходу на 1-2 ссылки. Информационный портал для специалистов в области надежности не является исключением и должен соответствовать всем современным требованиям [4].

Информационный портал для специалистов в области электронных средств предназначен для автоматизирования процесса расчета надежности современной радиоэлектронной аппаратуры и представляет собой информационно-справочную базу по характеристикам надежности компонентов компьютерной техники (ККТ) [5-7] и изделий электронной техники (ИЭТ) [8, 9]. Пользователь на портале может выполнить следующие действия:

1) выполнить поиск компонента без использования функций поиска, если он наверняка знает, что компонент имеется в базе данных (БД).

2) Выполнить поиск компонента в БД, используя функции поиска информационного портала

3) Выполнить сортировку найденных компонентов с использованием специального фильтра

4) Добавить новый компонент в БД портала

На рис. 1 показан примерный алгоритм взаимодействия пользователя с порталом при поиске известного ИЭТ в базе данных.

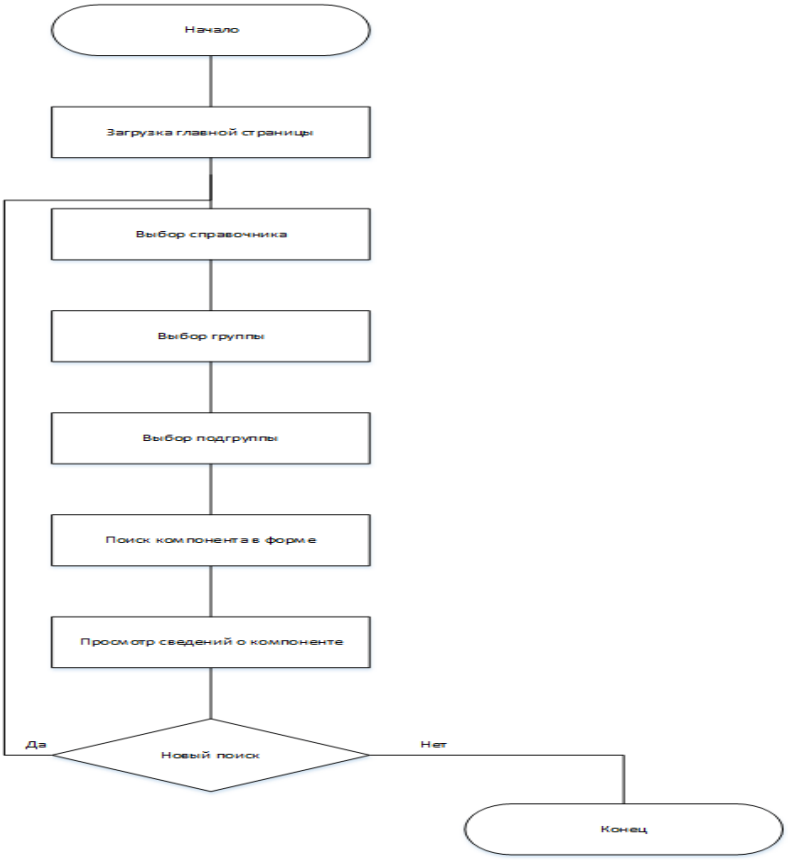

Рис. 1. Алгоритм поиска известного компонента

Пользователь загружает главную страницу информационного портала и в меню проводит выбор справочника. Помимо информации о надежности современных ИЭТ справочник содержит справочники по механическим компонентам и справочник по надежности компонентов компьютерной техники. После выбора справочника происходит выбор группы (например, конденсаторы) и подгруппы (напр. электролитические конденсаторы). После выбора компонента происходит открытие списка компонентов этой подгруппы, которые находятся в базе данных. Пользователь выбирает компонент из списка и просматривает его параметры. После просмотра он может начать новый поиск или завершить поиск.

На рис. 2 показан алгоритм взаимодействия пользователя с порталом при использовании функции поиска по БД.

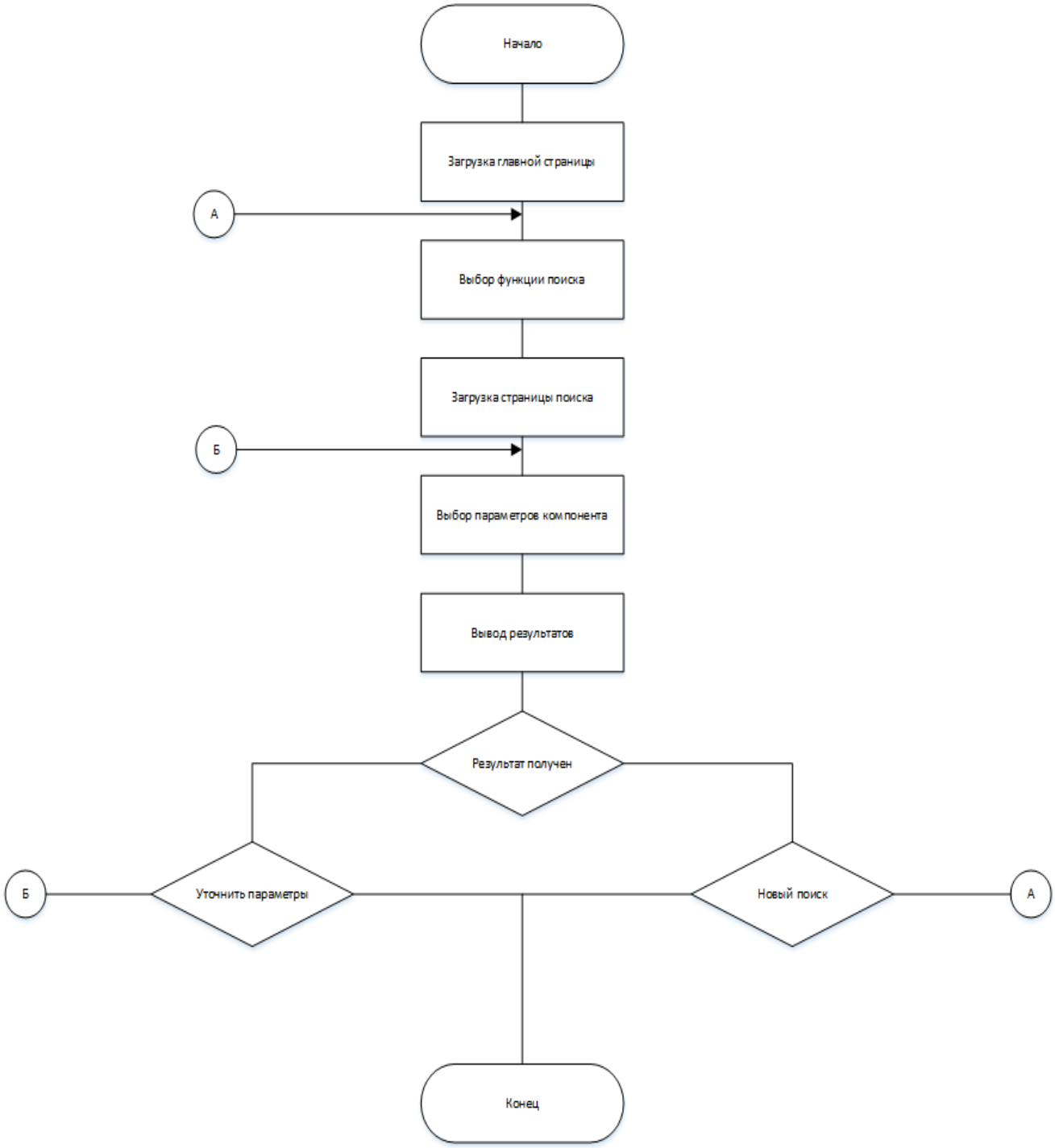

Рис. 2. Алгоритм поиска неизвестного компонента

Пользователь загружает главную страницу и выбирает функцию поиска компонента по базе данных. Происходит загрузка соответствующей страницы поиска, на которой пользователь заполняет форму поиска. После обработки запроса происходит вывод результатов в виде списка в том случае, если компонент найден. Если ничего не найдено система выдает соответствующее сообщение. Если пользователь удовлетворен результатом, то он может начать новый поиск или завершить работу с системой поиска. Если пользователь не получил результатов поиска, то он может уточнить запрос и повторить поиск [10].

При сортировке найденных компонентов и при просмотре любой подгруппы пользователь может воспользоваться специальным фильтром (рис. 3).

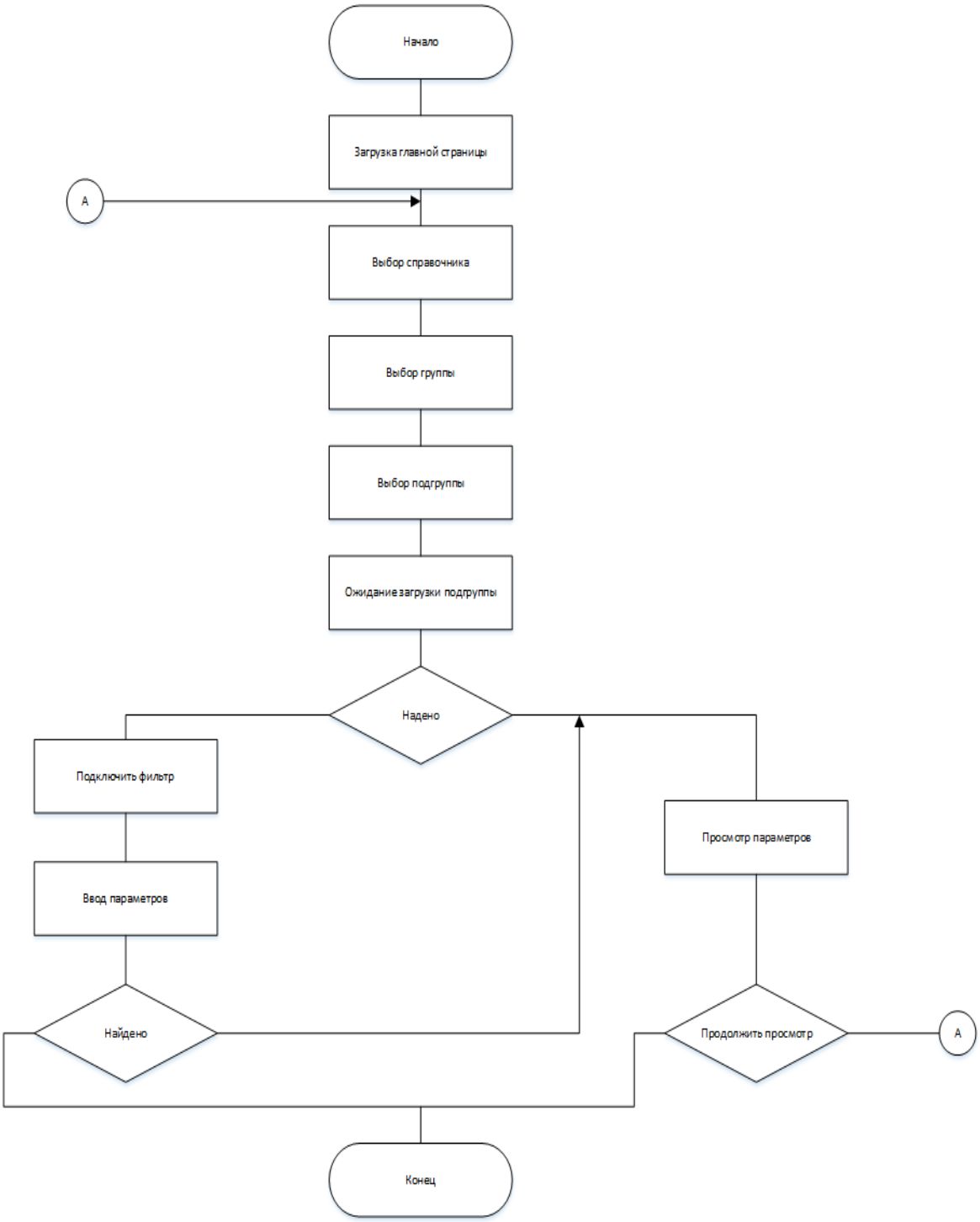

Рис. 3. Алгоритм использования фильтра

Пользователь загружает главную страницу и выбирает необходимый справочник. После выбора справочника происходит выбор группы и подгруппы. Если пользователь обнаружил необходимый элемент сразу, то он может его просмотреть не используя фильтр. Если же размер БД для этой подгруппы достаточно большой и содержит большое количество компонентов, то для облегчения поиска компонента можно воспользоваться функцией фильтра. Для этого необходимо его запустить и в появившейся форме указать необходимые параметры компонента. После подтверждения и обработки запроса пользователь сможет просмотреть те компоненты, которые удовлетворяют его требованиям. Если компонент не найден, то следует

уточнить поиск. Если же компонент после уточнения не найден, то, вероятно, в базе данных его нет и пользователь может добавить его самостоятельно.

Самостоятельное добавление компонента в базу данных информационного портала является его основным отличием от других порталов. Если пользователю известны параметры, необходимые для расчета надежности этого изделия и базовые параметры, то он может добавить его в БД. Информация попадет в базу данных после проверки администратором. Для добавления компонента необходимо заполнить соответвующую форму и отправить запрос. При удачном добавлении информации на рассмотрение администратора пользователь получит сообщение об успешном добавлении информации в БД портала.

Литература

1. Жаднов В.В. Информационные технологии в прогнозировании надежности электронных средств. / В.В. Жаднов // Информационные технологии в проектировании и производстве. № 1. 2012. – с. 20-25.

2. Жаднов В.В. Информационная технология обеспечения надежности сложных электронных средств военного и специального назначения. / В.В. Жаднов, Д.К. Авдеев, В.Н. Кулыгин и др. // Компоненты и технологии. - № 6. - 2011. - с. 168-174.

3. Абрамешин А.Е. Информационная технология обеспечения надежности электронных средств космических систем:научное издание. / А.Е. Абрамешин, В.В. Жаднов, С.Н. Полесский. / Отв. ред. В.В. Жаднов. – Екатеринбург: ООО «Форт Диалог-Исеть», 2012. – 565 с.

4. Дронов, В.А. HTML 5, CSS 3 и Web 2.0. Разработка современных Web-сайтов. / В.А. Дронов. – СПб: Издательство BHV-СПб, 2013. – 416 с.

5. Жаднов В.В. Оценка качества компонентов компьютерной техники. / В.В. Жаднов, С.Н. Полесский, С.Э. Якубов. // Надежность. № 3(26). 2008. – с. 26-35.

6. Жаднов, В.В. Прогнозирование качества ЭВС при проектировании. Учебное пособие. / В.В. Жаднов, С.Н. Полесский, С.Э. Якубов. – М.: ООО «СИНЦ», 2009. – 191 с.

7. Жаднов, В.В. Разработка информационно-справочной базы по характеристикам качества комплектующих электронных средств. / В.В. Жаднов, С.Н. Полесский, С.Э. Якубов. // Инновации в условиях развития информационно-телекоммуникационных технологий: М-лы научно-практической конференции. / Под ред. В.Г. Домрачева, С.У. Увайсова. Отв. за вып. А.В. Долматов, И.А. Иванов, Р.И. Увайсов. – М.: МИЭМ, 2008. – с. 111-113.

8. Цыганов, П.А. Информационный портал для специалистов в области надежности радиоэлектронных средств. / П.А. Цыганов, В.В. Жаднов. // Инновационные информационные технологии: Материалы международной научно-практической конференции. / Под ред. С.У. Увайсова; Отв. за вып. И.А. Иванов, Л.М. Агеева, Д.А. Дубоделова, В.Е. Еремина. – М.: МИЭМ, 2012. – с. 337-340.

9. Цыганов, П.А. Информационный портал для специалистов в области надежности радиоэлектронных средств. / П.А. Цыганов, В.В. Жаднов. // Современные проблемы радиоэлектроники: сб. науч. тр. / науч. ред. Г.Я. Шайдуров; отв. за вып. А.А. Левицкий. – Красноярск: Сиб. федер. ун-т, 2012. – с. 466-468.

10.Цыганов, П.А. Алгоритм работы программы поиска элементов по заданным параметрам в базе данных WEB-портала «НАДЕЖНОСТЬ ЭКБ». / П.А. Цыганов, В.В. Жаднов. // Современные проблемы радиоэлектроники: сб. науч. тр. / науч. ред. Г.Я. Шайдуров; отв. за вып. А.А. Левицкий. – Красноярск: Сиб. федер. ун-т, 2013. – с. 384-386.

Е.В. Шарапова
Муромский институт (филиал) Владимирского государственного университета
Россия, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Методы определения расстояния между строками

В современной научной среде актуальна задача сравнения различных источников информации. Сравнение обычно ведется с целью нахождения заимствований в текстах. Частичное или полное заимствование (плагиат) – серьезная проблема науки и образования, тормозящая развитие научной мысли и учебного процесса.

Электронные системы поиска заимствований в текстах призваны помочь в решении этой проблемы. Такие системы строятся на использовании различных алгоритмов сравнения текстов. Так как сравниваемые тексты (строки), скорее всего, будут разными, то имеет смысл рассмотреть алгоритмы, основанные на определении меры различия/схожести строк. Эта мера называется расстоянием между строками. Расстояние - это минимальное количество операций по изменению одного текстового символа, которое необходимо для превращения одной строки в другую. Для определения расстояния применяются алгоритм Хемминга и алгоритм Левенштейна.

1. Алгоритм Хемминга используется для нахождения расстояния между строками одинаковой длины, благодаря использованию операции «замена». Этот алгоритм не подходит для нахождения расстояния между различными по величине документами.

2. Алгоритм Левенштейна использует операции «замена», «вставка», «удаление», которые позволяют ему находить расстояние между разными по величине строками. Но время вычисления расстояния между строками непропорционально растет с увеличением размера сравниваемых строк. Поэтому применение этого алгоритма целесообразно только для сравнения нескольких страниц документов.

3. Алгоритм Дамерау-Левенштейна. Эта модификация расстояния Левенштейна учитывает еще и операцию «перестановка» («транспозиция») двух ближайших букв.

Метод вычисления расстояния Левенштейна лег в основу нескольких способов сравнения – алгоритмов поиска общих подпоследовательностей:

1. Алгоритм Вагнера и Фишера [1].

Этот метод основан на вычислении расстояния Левенштейна между префиксами строк (подстроками). Матрица редакционного предписания в этом алгоритме – итоговое представление расстояния Левенштейна (минимальные веса операций по изменению символов).

Число операций сравнения строк: k·p·b, где p и b – сравниваемые префиксы строк. Размер матрицы редакционных предписаний: (p+1)·(b+1).

		к	а	р	т	и	н	а
	0	1	2	3	4	5	6	7
к	1	0	1	2	3	4	5	6
о	2	1	1	2	3	4	5	6
р	3	2	2	1	2	3	4	5
з	4	3	3	2	2	3	4	5
и	5	4	4	3	3	2	3	4
н	6	5	5	4	4	3	2	3
а	7	6	5	4	4	3	2	2

Рис. 1. Пример матрицы предписаний.

Результаты работы алгоритма можно распространить на целые строки (принцип динамического программирования). Этот метод является наиболее простым способом представления редакционных предписаний.

2. Алгоритм Машека и Патерсона [2].

Этот алгоритм – модификация метода Вагнера и Фишера с применением подхода Арлазорова, Диница, Кронрода и Фараджева. В нем матрица расстояний разбивается на подматрицы, края которых вычисляются в соответствии с примыкающими к ним подматрицами.

Сложность алгоритма: $k \cdot (p \cdot b / \log(p))$.

Этот алгоритм быстрее алгоритма Вагнера и Фишера, но, по указанию самих его авторов [3], он достигает реальной скорости только при сравнении очень длинных строк.

3. Алгоритм Укконена [4].

Для реализации этого алгоритма требуется построение суффиксного дерева набора строк для минимизации времени поиска подстроки в строке.

Затрата памяти: $k \cdot l \cdot g$, где l – расстояние между строками, g – длина наименьшей строки.

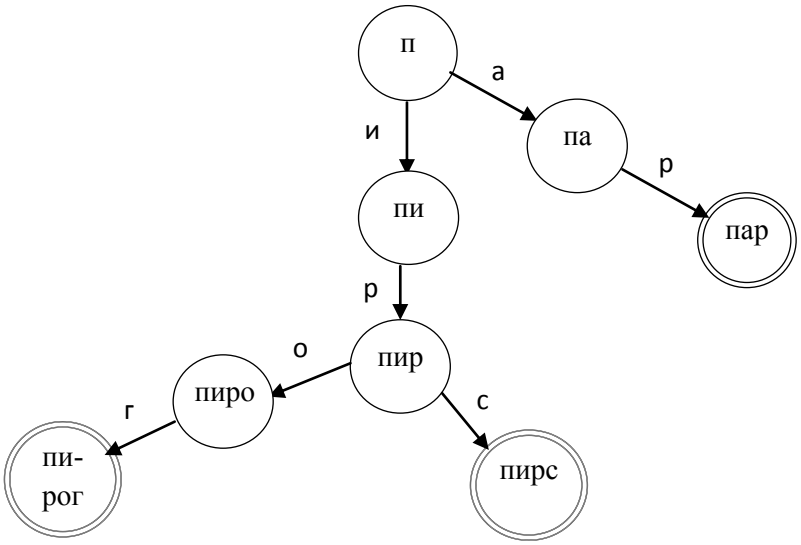

Рис.2. Пример суффиксного дерева для слов «Пирог», «Пирс» и «Пар».

Алгоритм Укконена более применим для поиска точных совпадений строк. При поиске сильно отличающихся текстов время его работы значительно увеличивается.

4. Алгоритм Хиршберга [5].

Алгоритм является модификацией алгоритма Вагнера и Фишера. В нем вычисляется не расстояние между строками, а расстояние между наиболее длинными общими подпоследовательностями.

Затраты памяти имеют линейную, а не квадратичную зависимость: $k \cdot (p+b)$.

Несмотря на то, что используемая память уменьшилась, количество времени сравнения по-прежнему квадратично.

5. Алгоритм Ханта и Шиманского [6].

Он основан на поиске максимального возрастающего пути на графе совпадений элементов сравниваемых строк. Время сравнения: $k \cdot (g+b) \cdot \log(b)$, где g – количество позиций, в которых символы строк совпадают. При определенных обстоятельствах сравнения этот алгоритм показывает хорошие результаты, но в других случаях время сравнения остается квадратичным.

Литература

1. Wagner R. A., Fischer M. J. The string-to-string correction problem // Journal ACM, Vol. 21, No. 1, 1974.
2. Masek W.J., Paterson M.S. A faster algorithm for computing string-edit distances // Journal of Computer and Systems Sciences, Vol. 20, No.1, 1980.
3. Masek W.J., Paterson M.S. How to compute string-edit distances quickly // In Sankofi D., Kruskall J.B. (eds.) Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading MA, 1983. Chapter 14, p. 337-49.
4. Гасфилд Дэн Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология — СПб.: Невский Диалект; БХВ-Петербург, 2003. – 654 с.
5. Hirschberg D. S. A linear space algorithm for computing maximal common subsequences // Communications of the ACM, Vol. 18, No. 6, 1975, p. 341–343.
6. Hunt J.W., Szymanski T.G. A fast algorithm for computing longest common subsequences // Communications of the ACM, Vol. 20, No. 5, 1977, p. 350-353.

Е.В. Шарапова
Муромский институт (филиал) Владимирского государственного университета
Россия, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Методы поиска точных совпадений в текстах

В современном мире людям доступно огромное количество информации. Каждый может найти то, что ему нужно в глобальной компьютерной сети Internet. Поиск можно осуществлять с разными целями: для организации досуга и отдыха, для профессиональной деятельности, для обучения и повышения квалификации. Но иногда поиск осуществляется с противоправной целью: присвоение чужого контента. Такое присвоение может производиться путем частичного изменения текста и интеграции его в другой текст, может остаться неизменным и быть смешанным с информационными блоками и так далее.

Компьютерные системы поиска заимствований в текстах призваны решить проблему плагиата. Такие системы строятся на различных алгоритмах сравнения текстов. Для начала рассмотрим алгоритмы поиска точных совпадений. Эти способы сравнения имеют цель нахождения абсолютного совпадения текстовых элементов.

1. Алгоритм Бойера-Мура [1].

При использовании этого метода поиска искомый текст обрабатывается с целью составления на его основе шаблонов сравнения и таблиц стоп-символов. Затем шаблон (подстрока) сравнивается с текстом. Сравнение подстрок происходит справа налево посимвольно, в соответствии с алфавитом строк. Если символы совпадают полностью, тогда искомый элемент найден, если нет, то шаблон сдвигается вправо на расстояние, зависящее от положения стоп-символов или суффиксов в тексте (см. рис.1).

Число действий, выполняемое этим алгоритмом для предварительной обработки, пропорционально произведению (p·s) длин строки поиска p и шаблона s. Наименьшее количество операций по алгоритму: (k·p/s), где k – коэффициент (для естественного языка k равен 0,2). Среднее число операций сравнения по алгоритму: k·(p+s).

Использование этого алгоритма удачно в случае, если поисковый текст предварительно не обрабатывался. Алгоритм требует составления таблиц стоп-символов и суффиксов шаблона. Это трудоемкие задачи, поэтому, в сочетании с медленной скоростью работы метода, они являются его недостатками.

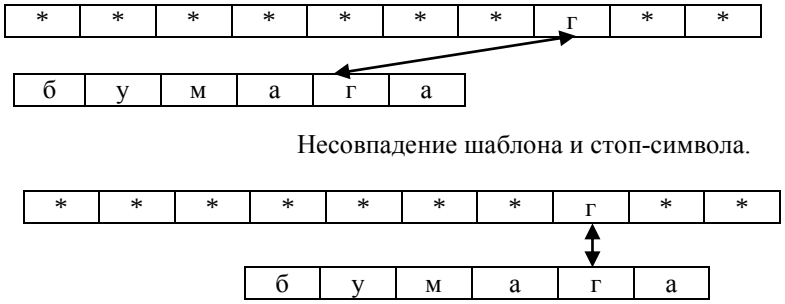

Рис.1. Действие алгоритма Бойера-Мура.

На основе алгоритма Бойера-Мура построены несколько других алгоритмов: алгоритм Хорспула, алгоритм Турбо-Бойера-Мура и др.

2. Алгоритм Кнута-Морриса-Пратта [2].

Этот алгоритм похож на алгоритм Бойера-Мура, но сравнение строк по нему производится слева направо. Если символы шаблона и элемента строки совпали, тогда найдено совпадение

текстовых элементов. Если не совпали, производится сдвиг шаблона вправо на позицию указателя, определяемую составленными заранее таблицами сдвига.

Число сравнений для предварительной обработки - k·p·s. Среднее количество операций сравнения по алгоритму: k·(p+s). Сильные и слабые стороны этого метода сравнения аналогичны характеристикам алгоритма Бойера-Мура.

Метод Кнута-Морриса-Пратта послужил основой для создания следующих алгоритмов: Colussi, Карпа-Рабина и др.

Литература

1. Boyer R. S., Moore J. S. A fast string searching algorithm // Communications of the ACM, Vol. 20, 1977. P. 762-772.

2. Knuth D.E., Morris J.H., Pratt V.R. Fast pattern matching in strings // SIAM Journal on Computing, Vol. 6, №1, 1977.