The approach to solving the TVVE mathematical model identification problem with the further use of this model for the fuel consumption optimization is offered. The problem solving is shown by an example of linearly-dynamic TVVE model with application of genetic algorithm for fine tuning of the model factors.

Key words: genetic algorithm, TVVE, mathematical model, control algorithm.

УДК 004.942, 001.53

Р. А. Файзрахманов, доктор экономических наук, профессор, Пермский государственный технический университет
Е. В. Долгова, доктор экономических наук, доцент, Пермский государственный технический университет
Р. Р. Файзрахманов, аспирант, Пермский государственный технический университет

МОДЕЛИРОВАНИЕ ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ВЕБ-СТРАНИЦ И ИЗВЛЕЧЕНИЯ ВЕБ-ИНФОРМАЦИИ*

Приведено описание процессов автоматической обработки веб-страниц и извлечения информации. Предложены и формально описаны основные модели веб-страниц (геометрическая и логическая), основанные на визуальном представлении. Показаны преимущества модели визуального представления для автоматической обработки веб-страниц и извлечения информации.

Ключевые слова: автоматическая обработка веб-страниц, извлечение веб-информации, моделирование процессов, геометрическая модель, логическая модель.

Веб является огромным хранилищем информации. Он играет важную роль в бизнесе, политике, науке и в повседневной жизни. Основным элементом являются веб-страницы (ВС), которые представляют информацию в слабоструктурированной или неструктурированной форме, используя такие всемирно известные стандарты, как HTML и XHTML. Данная форма представления используется исключительно для задания визуального форматирования, и также является удобной формой для хранения и передачи по сети Интернет. Но она не предназначения для отображения семантики и типов данных, которые содержит ВС, что делает автоматическую обработку довольно сложной проблемой.

Таким образом, существует проблема автоматизированной обработки веб-страниц (ОВС), их понимания компьютером, а также автоматизированного извлечения информации (ИзИ). ОВС и ИзИ играют важную роль в таких областях знаний, как информационный поиск, глубинный анализ данных, веб-адаптация, а также веб-доступность [1]. Помимо эффективности используемых алгоритмов, важную роль в данных проблемах играет представление веб-страницы, на уровне которого происходит решение той или иной задачи.

Процессы автоматической обработки веб-страниц и извлечения информации из визуального представления

Под **ОВС** мы подразумеваем обработку визуального представления ВС и представления ее в форме, понятной для компьютера. Процесс ОВС представим состоящей из двух основных этапов: анализа веб-страницы (АВС), результатом которого является геометрическая модель (ГМ) ВС, и понимания веб-страниц (ПВС), результатом которого является логическая модель (ЛМ) ВС – подобно обработке графических документов [2] (рис. 1).

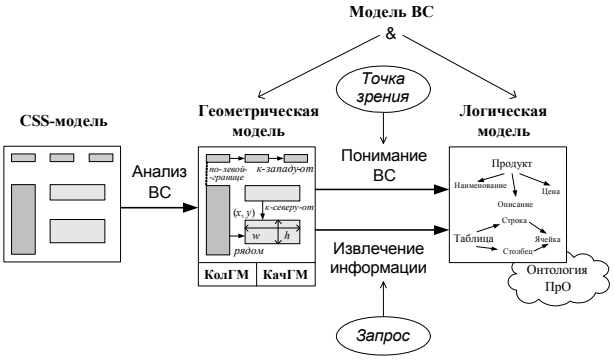

Рис. 1. Диаграмма, отражающая процесс обработки веб-страницы

Системы ИзИ предназначены для извлечения заведомо определенной информации согласно заданному запросу из некоторого набора источников и ее представления в виде, понятном компьютеру. Извлекаемая информация может иметь различные формы внутрисистемного представления, быть представлена в различных структурированных формах, например, в виде XML-документа, записей в базе данных или в виде утверждений в онтологии.

Существует 4 основные формы представления ВС, на основе которых может происходить ИзИ: 1) в виде исходного кода; 2) в виде дерева документа (DOM-дерево, или дерево тегов); 3) текстового представления и 4) визуального представления.

Получено 20.02.11
* При поддержке программы Евросоюза Erasmus Mundus External Cooperation Window.

В последнее время все больше внимания уделяется методам ИзИ из визуального представления (ИзИВП) ввиду того, что данное представление наиболее полно и точно отражает тот объем информации, который автор ВС планировал передать пользователю, где понятия и связи между ними выражены графически (в двумерном пространстве). Именно данное представление анализируется человеком при просмотре ВС. Остальные формы представления отражают лишь часть информации.

В частности, исходный код и дерево документа 1) не отражают графических аспектов ВС, хотя и широко применяются в задачах ИзИ ввиду их простоты и определенной структуры, которая является стандартом W3C; 2) нагружены дополнительными элементами не отражающих семантики определяемой автором ВС; 3) очень часто изменяются и делают недействительным применяемый алгоритм извлечения (враппер). Текстовое представление, которое может быть получено, например, при анализе исходного кода или благодаря текстовым веб-браузерам, предоставляет только текстовую составляющую веб-страницы, не отражая элементов визуального форматирования, а также не являясь монологическим текстом, и представляет собой огромную проблему, имеющую отношение к области обработки естественных языков.

Анализ визуального представления позволяет создавать системы ИзИ, которые не зависят от исходного кода ВС и применимы на значительно большем множестве ВС, чем существующие методы.

Таким образом, авторы представляют проблему ИзИВП как проблему ИзИ из ГМ и ее представление в виде ЛМ (см. рис. 1). Данная ЛМ в определенном смысле является подмножеством информации, полученной в процессе ОВС с соответствующим уровнем детализации. Можно сказать, что системы ИзИ работают интенсивным образом, в то время как системы ОВС – экстенсивным.

Модель веб-страницы. Формальное определение

Как было сказано выше, в процессе АВС, анализируется CSS-модель ВС, на основе которой формируется ГМ. На этапе ПВС или ИзИВП анализируется ГМ и на ее основе формируется ЛМ. Дадим основные определения.

Визуальное представление ВС (CSS-модель) формируется движком браузера согласно модели визуального форматирования CSS, основным структурным элементом которой является CSS-рамка [3]. Данное представление отображается в окне просмотра браузера и расположено в положительной полуплоскости, причем ордината направлена вниз.

Модель ВС M, предлагаемая в данной работе (см. рис. 1), это двойка, состоящая из геометрической (ГМ) G и логической (ГМ) L моделей: $M = \langle G, L \rangle$.

Геометрическая модель (ГМ) G ВС описывает ее визуальные характеристики, свойства и отношения между визуализированными элементами. Основным структурным элементом ГМ является геометрический объект (ГО), который представлен в виде прямоугольника. Можно выделить следующие типы ГО: элементарный ГО (ЭГО) и композитный ГО (КГО). ЭГО – прямоугольная область, обрамляющая содержимое веб-страницы, ему может быть поставлена в соответствие CSS-рамка. КГО – прямоугольная область, содержащая либо один, либо несколько ЭГО, либо другие КГО.

ГМ может быть основана как на количественной, так и на качественной[1] информации и представлять собой, соответственно, либо количественную (КолГМ), либо качественную (КачГМ) геометрические модели.

КолГМ представляет собой тройку $G_{qnt} = \langle \Im, A_{qnt}, B_{qnt} \rangle$. $\Im = \langle \Theta, L, \lambda \rangle$ определяет набор ГО ВС вместе с их метками. $\Theta = \{\theta_i\}$ – множество ГО; L – множество меток (например, HTML-тегов); $\lambda : \Theta \to L$ каждому ГО $\theta_i \in \Theta$ ставит в соответствие метку $\lambda(\theta_i) \in L$. $A_{qnt} = \{\alpha_i^{qnt}\}$, где $\alpha_i^{qnt} : \Theta \to range(\alpha_i^{qnt})$, иначе говоря $\alpha_i^{qnt}(\theta_i)$ – атрибут ГО θ_i, $range(\alpha_i^{qnt})$ – область принимаемых значений. $B_{qnt} = \{\beta_i^{qnt}\}$, где $\beta_i^{qnt} : \Theta \times \Theta \to range(\beta_i^{qnt})$, что можно интерпретировать, как отношение между ГО, выраженное количественной характеристикой.

ГО в КолГМ задается координатами левой верхней т.(x^-, y^-) и правой нижней т.(x^+, y^+), где $x, y \geq 0$, $x^+ > x^-$, $y^+ > y^-$. Следуя из определения, площадь ГО есть величина положительная.

Основными характеристиками (атрибутами) ГО θ_i КолГМ являются ширина $\alpha_w^{qnt}(\theta_i)$ и высота $\alpha_h^{qnt}(\theta_i)$. Основными характеристиками ЭГО являются атрибуты фона (цвет фона, фоновый рисунок при его наличии), границы (ширина, цвет, стиль) обрамляемого текста (размер, цвет, стиль), а также порядок рисования, используемый для определения видимости пересекающихся объектов и определяемый согласно правилам формирования стековых контекстов [3]. Между ГО могут быть количественно определены отношения дистанции $\beta_{dis}^{qnt}(\theta_i, \theta_j)$ и направления $\beta_{dir}^{qnt}(\theta_i, \theta_j)$. Схематично КолГМ представлена на рис. 2.

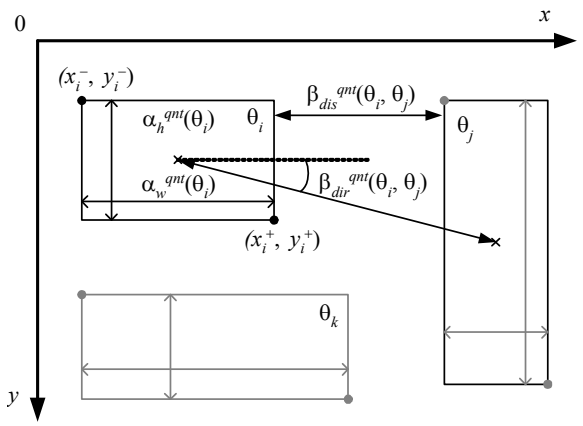

Рис. 2. Пример КолГМ

КачГМ строится на основе КолГМ и представляет собой тройку $G_{qlt} = \langle \Im, \Phi_u, \Phi_b \rangle$. $\Im$ совпадает с $\Im$ из

[1] Понятие «качественный» широко используется в области пространственного познания, например в работе [14], и также применяется в определении качественной геометрической модели, представленной в данной работе.

КолГМ G_{qnt}. $\Phi_u = \langle A_{qlt}, V, \iota, \alpha_{qlt} \rangle$ определяет свойства ГО $\theta_i \in \Theta$. $A_{ql} = \{a_i\}$ – множество свойств ГО; $V = \{v_i\}$ – множество значений, качественно выражающих свойства ГО; $\iota : A_{qlt} \to 2^V$ определяет область значений для каждого свойства $a_i \in A_{qlt}$; $\alpha_{qlt} : \Theta \to 2^V$ определяет значения свойств для каждого ГО и является одноместной (унарной) функцией. $\Phi_b = \langle B_{qlt}, R, \tau, \beta_{qlt} \rangle$ определяет отношения на множестве ГО Θ. $B_{qlt} = \{b_i\}$ – множество различных типов отношений; $R = \{r_i\}$ – множество отношений, определенных на Θ; $\tau : B_{qlt} \to 2^R$; $\beta_{qlt} : \Theta \times \Theta \to 2^R$ определяет бинарные отношения между ГО.

Элементы $a_i \in A_{qlt}$ и $b_j \in B_{qlt}$ можно интерпретировать как лингвистические переменные, а $\iota(a_i)$ и $\tau(b_j)$, соответственно, как множества принимаемых значений. Для каждого значения свойства и для каждого отношения должны быть определены, соответственно, ровно одно свойство и один тип отношения, то есть

$$\forall (a_i, a_j \in A_{qlt}) : a_i \neq a_j \to \iota(a_i) \cap \iota(a_j) = \varnothing,$$

$$\forall (b_i, b_j \in B_{qlt}) : b_i \neq b_j \to \tau(b_i) \cap \tau(b_j) = \varnothing.$$

В случае если ГО имеет более одного значения для одного и того же свойства (1) или более одного отношения одного и того же типа (2), речь идет о наличии неопределенности в КачГМ.

$$\exists (\theta_i \in \Theta; v_i, v_j \in \alpha_{qlt}(\theta_i)) : v_i \neq v_j \wedge \iota^{-1}(v_i) = \iota^{-1}(v_j), \quad (1)$$

$$\exists (\theta_i, \theta_j \in \Theta; r_i, r_j \in \beta_{qlt}(\theta_i, \theta_j)) : r_i \neq r_j \wedge \tau^{-1}(r_i) = \tau^{-1}(r_j). \quad (2)$$

Основными характеристиками ГО КачГМ являются ширина и высота, имеющие такие значения, как «очень маленький», «маленький», «средний», «большой», «очень большой». Основными характеристиками ЭГО КачГМ являются атрибуты фона, границы обрамляемого текста, выраженные качественными значениями.

Отношения между ГО КачГМ могут быть различной природы, сравнивая их, например, по яркости («ярче», «темнее»), размеру («больше», «меньше») или любым другим признакам (например «более броский»). Но не менее важным является пространственная ориентация и пространственные отношения между объектами, позволяющие нам определять, к примеру, навигационное меню как набор текстовых блоков, ориентированных горизонтально или же вертикально.

Пространственные отношения, определяемые на множестве ГО, можно разделить на следующие группы: *топологические* («касается», «внутри», «перекрывает», «не-имеет-общих-точек»), *направления* («к-северу-от», «к-северо-востоку-от», «к-востоку-от» и т.д.), *дистанции* («очень-близко», «близко», «не-далеко», «далеко», «очень-далеко»), и *отношения выравнивания* («по-левому-краю», «по-правому-краю», «центрировано-вертикально» и также для вертикального выравнивания) (рис. 3). Количество значений для топологической группы отношений, направления и дистанции может отличаться.

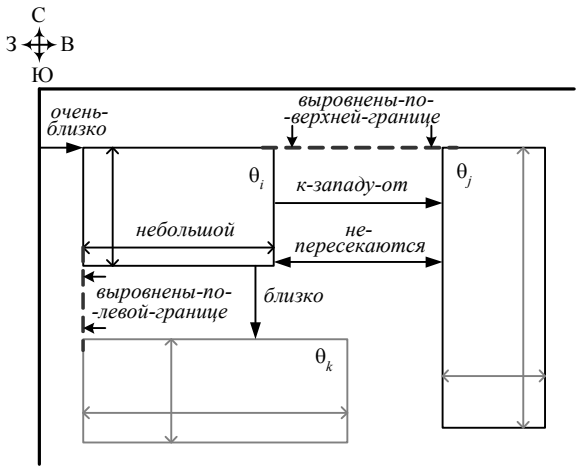

Рис. 3. Пример КачГМ

Преимущество использования качественной информации перед количественной определяется возможностью создания более робастных алгоритмов и методов, например с использованием нечеткой логики, не зависящих от конкретных числовых данных. Нужно заметить, что именно пространственные отношения и качественные характеристики ГО, такие как размер, цвет, анализируются человеком при понимании веб-страницы. Таким образом, использование данных отношений позволяет создавать алгоритмы, устойчивые на большом множестве ВС.

Логическая модель L (ЛМ) представляет собой тройку $L = \langle \Theta, O, \nu \rangle$, где Θ – множество ГО, O – онтология предметной области (ПрО), $\nu : \Theta \to |O|$ ставит функциональное соответствие между множеством ГО и понятиями в онтологии O.

Связывая множество ГО с понятиями определяемыми в онтологии ПрО, ЛМ является интерпретацией ГМ или же ее части. Данная информация может далее использоваться для автоматической обработки аналитическими интеллектуальными системами, основанными на знаниях. Использование онтологии не только делает ВС понятной для компьютера, но и делает ее составляющей семантического Веба, так как определяет ее метаинформацию в случае использования широко известных стандартов, таких как RDF и OWL.

Заключение

В работе рассмотрены процессы автоматической обработки веб-страниц и извлечения информации, описана предложенная авторами модель представления ВС, основанная на ее визуальном представлении, в которой учтены основные визуальные характеристики в форме, удобной для выполнения задач ОВС и ИзИ, а также представления добытой информации. Данные модели позволят создавать эффективные методы, обеспечивая более детальное понимание веб-страницы и устойчивые алгоритмы ИзИ, применимые на большем множестве ВС. Это позволяет, в свою очередь, создавать информационно-поисковые

системы, имеющие большую точность и полноту в отыскании релевантной информации, и услужливые технологии, позволяющие сделать Веб более доступным.

Список литературы

1. A unified ontology-based web page model for improving accessibility / R. R. Fayzrakhmanov [et al.] // Proc. of the 19th international conference on World Wide Web. – 2010. – P. 1087–1088.

2. *Haralick R. M.* Document image understanding: geometric and logical layout // Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. – 1994. – P. 385–390.

3. Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification / B. Bos [et al.] // World Wide Web Consortium. 2009. – URL: http://www.w3.org/TR/ 2009/CR-CSS2-20090908/ (дата обращения: 20.01.2011).

R. A. Fayzrakhmanov, Doctor of Economics, Professor, Perm State Technical University
E. V. Dolgova, Doctor of Economics, Associate Professor, Perm State Technical University
R. R. Fayzrakhmanov, Postgraduate Student, Perm State Technical University

Modeling of Information Representation in the Tasks of Web Page Processing and Web Information Extraction

The Web page processing and Web information extraction are described. The main Web page models based on visual representation, such as geometrical and logical ones, are introduced and formally described. Advantages of the models for Web page processing and Web information extraction are shown.

Key words: Web page processing, Web information extraction, process modeling, geometrical model, logical model.

УДК 621.512.011.56

Ю. Ф. Рубцов, кандидат технических наук, Пермский государственный технический университет
Д. Ю. Рубцов, ОАО «Научно-исследовательский институт управляющих машин и систем», Пермь

ПРЕДЕЛЬНО ДОПУСТИМЫЕ ПОГРЕШНОСТИ ИЗМЕРИТЕЛЬНЫХ КАНАЛОВ АВТОМАТИЗИРОВАННОГО РАБОЧЕГО МЕСТА ИСПЫТАНИЙ ДВИГАТЕЛЯ ПОСТОЯННОГО ТОКА ЭДУ-133

Рассматривается методика расчета предельно допустимых погрешностей измерительных каналов автоматизированного рабочего места испытаний тягового двигателя ЭДУ-133. Представлены исходные данные для расчета предельно допустимых погрешностей при испытаниях тягового двигателя постоянного тока ЭДУ-133.

Ключевые слова: методика, погрешность, параметры, компонент, канал, измерение.

Автоматизированное рабочее место испытаний режимных параметров (АРМИ-РП) тягового двигателя постоянного тока ЭДУ-133 предназначено для проведения приемочных, периодических и приемо-сдаточных испытаний, для обработки и хранения результатов испытаний. Разные задачи испытаний требуют различной точности, что, в свою очередь, делает задачу определения предельно допустимых погрешностей измерительных каналов при проведении испытаний электрических двигателей постоянного тока актуальной.

Необходимую точность измерения устанавливаем, нормируя значение предельно допустимой погрешности измерения [1]. Принцип нормирования состоит в том, чтобы предельно допустимая погрешность не оказывала существенного влияния на достоверность результатов измерений при испытаниях. Характеристики погрешностей измерительных каналов (ИК) автоматизированного рабочего места испытаний режимных и других испытательных параметров (АРМИ-РП) нормируются путем установления предела допускаемой относительной погрешности ИК в предусмотренных рабочих условиях применения при доверительной вероятности 0,95. Допустимые погрешности измерения испытательных параметров испытуемого двигателя (ИД) ЭДУ-133, входящие в состав АРМИ-РП, должны удовлетворять требованиям, указанным в табл. 1 и 2. Метрологические характеристики (МХ) АРМИ-РП нормируются без выделения составляющих (основной и дополнительной погрешности ИК-параметров).

Таблица 1. **Предельно допустимые погрешности ИК-параметров с прямым измерением**

Наименование класса параметров	Погрешность, %
Напряжение постоянного тока	0,3/0,8
Постоянный ток	0,5/1,0
Сопротивления изоляции	2,5
Температура	2,0
Частота вращения	1,0
Виброскорость	6,0

Получено 28.02.11