реализующего модель многослойного персептрона, используемого для распознавания голосовых команд. Применение голосового управления позволит сократить время и стоимость проектирования на 15-20%.

Практическая ценность работы заключается в уменьшении времени и стоимости технологической подготовки роботизированного производства за счёт голосового задания управляющих сигналов.

В дальнейшем планируется интегрировать разработанное программное обеспечение в систему управления роботами MR-999e и PM-01, также реализовать голосовое управление при помощи других моделей представления ИНС.

Литература: 1. *Искусственный* интеллект: Применение в интегрированных производственных системах / Под ред. Э. Кьюсиака: Пер. с англ. А.П.Фомина /Под ред. А.И. Дащенко, Е.В. Левнера. М.: Машиностроение, 1991. 544с. **2.** *Рабинер Л., Гоулд Б.* Теория и применение цифровой обработки сигналов. М.: Мир, 1978. **3.** *Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи* и др. Распознавание речи // Прикладные нечёткие системы /Под ред. Тэрано Т., Асаи К., Сугено М.. М.: Мир, 1993. **4.** *Комарцова Л.Г., Максимов А.В.* Нейрокомпьютеры. М.:Издательство МГТУ им. Н.Э.Баумана, 2002. 320с. **5.** *Терехов С.А.* Лекции по теории и приложениям нейронных сетей. 1994. Лаборатория Искусственных Нейронных Сетей НТО-2, ВНИИТФ, Снежинск **6.** *Нейроинформатика* / А. Н. Горбань, В. Л. Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов и др. Новосибирск: Наука, 1998. 296 С. **7.** *Головко В. А.* Нейронные сети: обучение, организация и применение. М.: ИПРЖР, 2001.

Поступила в редколлегию 14.02.2007

Рецензент: д-р техн. наук, проф. Ильченко Б.С.

Невлюдов Игорь Шакирович, д-р техн. наук, проф. ХНУРЭ. Научные интересы: технология приборостроения, гибкие производственные системы, робототехника. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86.

Цымбал Александр Михайлович, канд. техн. наук, доцент, докторант ХНУРЭ. Научные интересы: системы программирования, системы искусственного интеллекта. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86, e-mail: mcdulcimer@kture.kharkov.ua.

Милютина Светлана Святославовна, аспирантка кафедры ТАПР ХНУРЭ. Научные интересы: системы программирования, системы искусственного интеллекта. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86.

УДК519.7

ОБЪЕКТНОЕ ПРЕДСТАВЛЕНИЕ ЭЛЕКТРОННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

ГВОЗДИНСКИЙ А.Н., ГУБИН В.А.

Рассматривается проблема формализации содержимого электронных текстовых документов. Документы представляются в виде совокупности объектов двух видов – объектов контейнеров и атомарных объектов. Каждая группа объектов отражает либо топологический, либо информационный аспект содержимого документа.

Актуальность исследования

Бурное развитие вычислительной техники, сети Internet, приход компьютеров практически в каждый офис, в каждый дом порождает тенденцию увеличения удельного веса представления информации в электронном виде. С развитием концепции электронного документооборота на первый план выходят электронные документы как носители и источники информации, а документы на бумаге отходят на второй план, уступая свои позиции особенно в тех областях, где требуется высокий уровень мобильности и оперативности.

С другой стороны, бурное развитие сети Internet и ее общедоступность сделали практически неограниченным доступный информационный массив. Большая часть этого массива изначально не предполагала возможность автоматизированной обработки. Это породило необходимость перехода от методов обработки документов на бумажных носителях к развитию и совершенствованию технологий автоматизированной обработки электронных источников информации.

Данные обстоятельства привели к возникновению и развитию технологии Text Mining – современного направления интеллектуального анализа и обработки текстовых данных. Эта технология, являясь одним из направлений Data Mining, позволяет решать разнообразные задачи, возникающие при анализе больших электронных массивов неструктурированной информации.

Отличительной особенностью современных подходов в Text Mining является то, что единицей анализа содержимого электронных текстовых документов есть слово. При этом игнорируется то обстоятельство, что документы определенного класса могут состоять из текстовых фрагментов, обособленных относительно других фрагментов и представляющих ценность как некоторая неделимая единица. Для определенного класса задач, в частности, для задач идентификации данных в текстовых документах, это может быть достаточно существенным недостатком. Настоящая работа предлагает подход, устраняющий этот недостаток.

Целью исследования является формализация содержимого электронных текстовых документов [1]. При этом документы представляются в виде совокупности объектов двух видов – объектов контейнеров и атомарных объектов. Первая группа объектов отражает топологию документа, вторая – его информационное содержимое. Также важно, чтобы о каждом обособленном текстовом фрагменте документа сохранялась информация о контексте его появления.

Задачи исследования: разработка спецификации объектов контейнеров и атомарных объектов; разработка методики определения того, какие фрагменты исходного документа необходимо отнести к объектам того

или иного типа и какие отношения между этими объектами могут быть установлены.

Объекты каждой группы характеризуются совокупностью свойств, значения которых отражают особенности конкретного объекта и его отношения с другими объектами. При этом каждому обособленному текстовому фрагменту документа соответствует атомарный объект, ключевым свойством которого является значение соответствующей текстовой строки.

Модель объектного представления документов

Пусть имеется исходное пространство электронных текстовых документов Ω, содержащее документы $D_1, D_2, \ldots, D_N$. В этом случае Ω можно интерпретировать как множество, содержащее элементы $D_1, D_2, \ldots, D_N$, где N – количество документов в пространстве Ω. Таким образом: $\Omega = \{ D_1, D_2, \ldots, D_N\}$.

Предполагаем, что документы, входящие в это пространство, обладают структурой, т.е. существует некоторая внутренняя разметка документа. Данное обстоятельство позволяет представить документ как набор образующих его элементов, которые могут иметь те или иные свойства, отличающие их от других элементов. К элементам можно отнести абзацы, таблицы, нумерованные и ненумерованные списки и т.п. Примерами документов, обладающих внутренней структурной разметкой, могут быть документы, представленные в формате HTML, DOC, RTF и в других аналогичных форматах.

Если каждый элемент документа или часть элемента интерпретировать как объект, то документ можно представить в виде неупорядоченного множества объектов: $D_i = \{\theta_1, \theta_2, \ldots, \theta_{n_i}\}$, i = 1, … ,N, где n_i – количество объектов в i -м документе.

Необходимо добиться того, чтобы данное разбиение отражало и топологию, и содержимое документа. Для этого вводятся два типа объектов: объекты-контейнеры и атомарные объекты. К объектам-контейнерам отнесем сам документ, абзац, таблицу, ее строку и ячейку, нумерованный и ненумерованный список, элемент списка и т.п. К атомарным объектам отнесем содержимое абзаца, выделенную тем или иным способом часть содержимого абзаца, содержимое ячейки таблицы, содержимое элемента списка и т.п. При этом предполагается, что содержимое атомарных объектов не может быть пустым или подвергнуто дальнейшему разбиению. Из такого определения объектов-контейнеров и атомарных объектов следует, что объекты-контейнеры могут содержать один или более других объектов-контейнеров или один или более атомарных объектов.

Обозначим объекты-контейнеры как φ и атомарные объекты как ψ. В этом случае каждый документ пространства Ω может быть представлен в следующем виде:

$$D_i = \{\varphi_1, \varphi_2, \ldots, \varphi_{p_i}, \psi_1, \psi_2, \ldots, \psi_{l_i}\}, \; i = 1, \ldots, N,$$

где p_i – количество объектов-контейнеров, а l_i – количество атомарных объектов в i-м документе, или

$$D_i = \{\Phi_i, \Psi_i\}, \; i = 1, \ldots, N,$$

где $\Phi_i = \{\varphi_1, \varphi_2, \ldots, \varphi_{p_i}\}$, а $\Psi_i = \{\psi_1, \psi_2, \ldots, \psi_{l_i}\}$.

При этом объекты-контейнеры могут находиться между собой в отношении владения или следования. Отношение следования между объектами-контейнерами отражает взаимное расположение различных элементов в документе. Отношение владения отражает вложенность одних элементов в другие. Например, список может включать в качестве своего элемента другой список (случай вложенных списков), таблица состоит из строк и т.п. Особенностью объектов-контейнеров является отсутствие в них текстового содержимого. Таким образом, совокупность объектов-контейнеров $\Phi_i = \{\varphi_1, \varphi_2, \ldots, \varphi_{p_i}\}$ отражает топологию документа.

Ключевой особенностью атомарных объектов является наличие у них в качестве одного из свойств текстовой строки, являющейся структурно-обособленным фрагментом текстового содержимого документа. Атомарные объекты могут находиться между собой в отношении ассоциации. Например, обособленные фрагменты одного и того же абзаца, элемент списка более высокого уровня с каждым простым элементом вложенного списка, содержимое следующих друг за другом абзацев и т.п. Таким образом, совокупность атомарных объектов $\Psi_i = \{\psi_1, \psi_2, \ldots, \psi_{l_i}\}$ отражает текстовое содержимое фрагмента.

Предполагается, что контекст данных и их значение в анализируемых документах соответствуют атомарным объектам, находящимся между собой в отношении ассоциации.

Исходя из логики определения объектов-контейнеров и атомарных объектов, можно сделать вывод, что между собой эти объекты могут находиться только в отношении владения.

Схематически сценарий преобразования представлен на рис. 1.

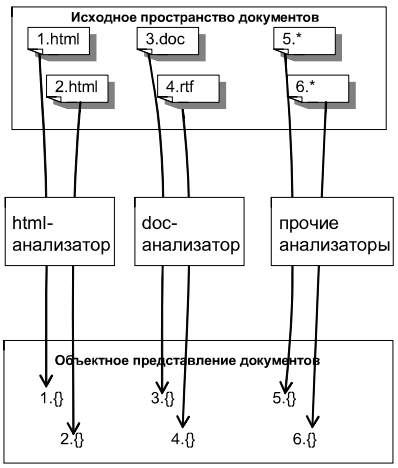

Рис. 1. Преобразование с использованием специализированных анализаторов

В идеале, для каждого существующего формата представления электронных текстовых документов необходимо разработать свой оригинальный анализатор. Но, учитывая, что практически все популярные форматы снабжены инструментальными средствами преобразования соответствующих документов в формат HTML, достаточно разработать HTML-анализатор. При этом схема преобразования может выглядеть так, как показано на рис. 2.

Необходимо, чтобы в HTML-анализаторе решались следующие задачи:

1. Первичная обработка исходного HTML-документа. В частности, исправление грамматических ошибок.

2. Представление топологии входного HTML-документа в виде совокупности объектов-контейнеров.

3. Представление текстового содержимого входного HTML-документа в виде совокупности атомарных объектов.

4. Установление отношений между объектами и идентификация их свойств.

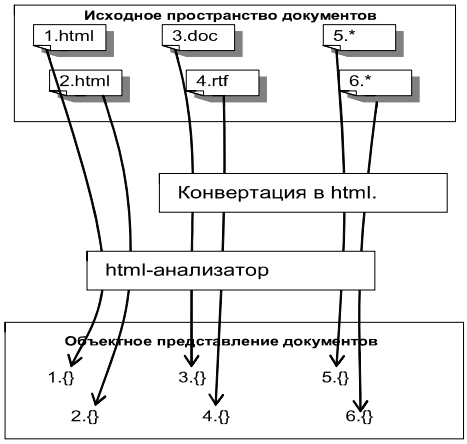

Рис. 2. Преобразование с использованием только HTML-анализатора

В процессе объектного представления необходимо идентифицировать ряд свойств объектов-контейнеров и атомарных объектов. У объектов контейнеров: тип контейнера, объект-владелец, предыдущий контейнер, следующий контейнер. У атомарных объектов: соответствующий текстовый фрагмент, форматирование, контейнер-владелец,

Выводы

Разработан оригинальный подход к формализации содержимого электронных текстовых документов. В основе этого подхода лежит объектный подход, а сами документы представлены как совокупности атомарных объектов и объектов контейнеров. Данная модель позволяет интерпретировать некоторые обособленные фрагменты документа как самостоятельные и неделимые единицы анализа.

Этот подход позволяет также получить некоторое универсальное представление для документов с различным исходным форматом и упрощает задачу идентификации данных, содержащихся в текстовых документах.

Научная новизна. Предложен подход к формализации содержимого электронных текстовых документов.

Практическая значимость. Использование предложенной в работе модели позволит существенно упростить задачу идентификации данных в электронных текстовых документах определенного класса [2], в частности, в документах, в которых объективно присутствуют обособленные контекст и значение данных, но отсутствуют формальные признаки, указывающие на то, что есть что.

Литературы: 1. *Гвоздинский А.Н., Губин В.О., Якимова Н.А.* О природе слабоструктурированных источников информации // Труды 10-й Международной научной конференции «Теория и техника передачи, приема и обработки информации». Туапсе, 2004. С. 68-69. **2.** *Гвоздинский А.Н., Губин В.О., Якимова Н.А.* О проблеме поиска информации в слабоструктурированных источниках / / Труды 11-й Международной научной конференции «Теория и техника передачи, приема и обработки информации». Туапсе, 2005. С. 72-73.

Поступила в редколлегию 20.01.2007

Рецензент: д-р техн. наук, проф. Соколов А.Ю.

Гвоздинский Анатолий Николаевич, канд. техн. наук, профессор кафедры искусственного интеллекта ХНУРЭ. Научные интересы: оптимизация процедур принятия решений в сложных системах управления. Адрес: Украина, 61166, Харьков, ул. акад. Ляпунова, 7, кв. 9, тел. 702-38-23.

Губин Вадим Александрович, преподаватель кафедры искусственного интеллекта ХНУРЭ. Научные интересы: интеллектуальный анализ текстовых данных. Адрес: Украина, 61053, Харьков, ул. Гвардейцев-Широнинцев, 23, кв. 286, тел. 710-64-12.