Белоусова Светлана Алексеевна – Южный федеральный университет; e-mail: s.belousova2011@gmail.com; 347928, г. Таганрог, Некрасовский, 44; тел.: 88634371787; кафедра системного анализа и телекоммуникаций; аспирант; ассистент.

Рогозов Юрий Иванович – e-mail: rogozov@tti.sfedu.ru; кафедра системного анализа и телекоммуникаций; зав. кафедрой; д.т.н.; профессор.

Belousova Svetlana Alexeevna – Southern Federal University; e-mail: s.belousova2011@gmail.com; 44, Nekrasovsky, Taganrog, 347928, Russia; phone: +78634371787; the department of system analysis and telecommunications; postgraduate student; assistant.

Rogozov Yury Ivanovich – e-mail: rogozov@tti.sfedu.ru; the department of system analysis and telecommunications; head the department; dr. of eng. sc.; professor.

УДК 004.912 + 004.822

В.В. Ланин

МНОГОАСПЕКТНАЯ ОНТОЛОГИЯ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ КАК ОСНОВА ФУНКЦИОНИРОВАНИЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ*

Предлагается подход к проектированию и организации функционирования информационных систем на основе обработки неструктурированной информации, представленной в электронных документах в различных форматах. Обработка документов основана на их семантическом индексировании и включении в них дополнительной метаинформации. Описана модель документа, позволяющая формализовать алгоритмы интеллектуальной обработки документов, предоставляющая широкие возможности для интеграции документов с онтологическими ресурсами. Семантическое индексирование основано на многоаспектной онтологии, описывающей структуру и семантику документа. Для обработки документов предлагается использовать агентный подход, который позволяет решить проблему включения бизнес-логики в документы. Система становится более гибкой, интеллектуальной, адаптируемой к изменению среды. Предлагаемый подход позволяет решить широкий спектр задач, связанных с использованием электронных документов в информационных системах на всех этапах их жизненного цикла, что позволяет говорить о документно-ориентированной парадигме поддержки всего жизненного цикла информационных систем.

Онтология; электронный документ; модель документа; информационная система.

V.V. Lanin

MULTIDIMENSIONAL ONTOLOGY OF ELECTRONIC DOCUMENT AS A BASE OF INFORATION SYSTEM

An approach to designing and organization of information systems operation based on the processing of unstructured information presented in different formats of electronic documents is proposed. Documents processing relies on its semantic indexing and inclusion in its additional meta-information. The paper presents a model of the document, allowing to formalize algorithms of intelligent processing documents and providing opportunities for integration document with ontological resources. Semantic indexing is based on a multidimensional ontology, which describes the structure and semantics of the document. This agent-based approach to semantic indexing of documents allows solving the problem of including business logic to documents. The system becomes more flexible, intelligent, adaptable to dynamic environments. The proposed ap-

* Исследование выполнено при финансовой поддержке РФФИ в рамках проекта № 14-07-31273-мол_а.

proach allows to solve a wide range of tasks associated with the use of electronic documents in the information system at all stages of their life cycle. That makes possible to use term suggesting that the "document-oriented paradigm" to support the whole life cycle of information system.

Ontology; electronic document; document model; information system.

Введение. В современных информационных системах (ИС) наблюдается переход от обработки структурированных данных к оперированию документами, неструктурированными данными. Важной частью информационного пространства стали новые классы систем (социальные сети, корпоративные порталы, wiki-ресурсы и т.д.). Ключевым для таких систем является «контент», это понятие можно обобщить как «электронный документ».

Современный подход к определению «*электронный документ»* кроме представления содержимого требует наличия *метаданных*, описывающих структуру и семантику представленных в документе данных. Благодаря такому подходу, обработка электронных документов может быть организована на качественно ином уровне, так как становится возможным автоматический интеллектуальный анализ информации. Эта концепция заложена в проект Semantic Web, однако состояние проекта «семантической паутины» в силу целого ряда причин ещё далеко от практической реализации. Однако идеи, заложенные в Semantic Web [2], могут быть реализованы в рамках отдельно взятой информационной системы благодаря меньшему масштабу ее предметной области [1]. В настоящее время данные, необходимые для обработки документов, рассредоточены (хранятся как в самом документе, так и в базах данных ИС, обрабатывающих документы) и специфичны для каждой из задач, решаемых в течение жизненного цикла документа в ИС. Появляется необходимость использования единого механизма представления информации о документе. Решением может стать онтологический ресурс, описывающий различные аспекты электронного документа, рассматриваемые в течение всего его жизненного цикла. Этот ресурс может стать основой для решения широкого спектра задач, связанных с обработкой электронных документов в ИС.

Для решения поставленных задач необходимо разработать *модель электронного документа*, позволяющую включить в него метаинформацию, и *онтологический ресурс*, являющийся базой для семантического индексирования содержимого документа, создать *механизм обработки документов*.

Модель документа. Электронный документ представляет собой *набор структурных элементов*, называемых в данной работе *фрагментами*. Примерами фрагментов могут служить: таблица, заголовок, реквизиты углового бланка и т.д. Таким образом, документ может быть представлен четверкой вида

$$d = (S(F, R), C, o, M).$$

где $S(F, R)$ – ориентированный гиперграф, вершинам которого сопоставлены элементы множества F (множество F – это множество фрагментов документа, а R – это множество ребер графа, соответствующее связям между фрагментами); элементы множества C представляют информационное содержание документа (его контент); o – онтология документа, M – отображение множества F на концепты онтологии o. Рассмотрим подробнее описанные компоненты.

Гиперграф $S(F, R)$ задает *отношение между фрагментами документа*. Ориентированность графа необходима, например, для отслеживания связей «часть-целое» между фрагментами. Вершины, входящие в ребро, пронумерованы, что позволяет установить порядок следования фрагментов в тексте документа. Очевидно, что ребро, включающее все вершины, соответствует документу целиком.

Фрагменты могут быть двух видов: элементарные и составные. *Элементарные* фрагменты представляют простейшие неделимые элементы, такие как заголовок или дата составления документа, а *составные* содержат в себе другие фрагменты.

Определим формально *фрагмент* как пару вида

$$f = (stat,\ inf),\quad inf = \begin{cases} F^*, F^* \subseteq F; \\ c, c \in C. \end{cases}$$

Здесь *stat* – это статическая часть фрагмента, она может быть представлена текстом, изображением, ссылкой, каким-либо специальным символом, кроме того, здесь может содержаться и информация для представления фрагмента; *inf* – это часть фрагмента, которая либо указывает место для размещения элемента содержания *c* ($c \in C$), либо содержит множество фрагментов F^*.

Традиционно для представления документа используются обычные графы, чаще всего деревья (например, формат XML). Древовидная структура описания значительно упрощает работу с документом, но вместе с тем вносит и существенные ограничения. Выбор гиперграфа для представления структуры документа обосновывается возможностями гиперграфов представлять произвольные связи между фрагментами документа и их множествами.

В описанных выше обозначениях *шаблон документа* можно определить как $t = (S(F, R), C_0)$, где C_0 – *первичный контент* (например, стандартные заголовки, включенные в шаблон, и т.д.).

Учитывая специфику решаемых в данной работе задач, конкретизируем понятие *онтологии*:

$$o = (C, R, A),$$

где C – множество *понятий* (*концептов*) онтологии, R – множество *отношений* между концептами, A – множество *аксиом*, заданных на онтологии. Концептами могут быть как классы, так и экземпляры этих классов, а аксиомы используются для задания ограничений и правил, которые не могут быть выражены через отношения.

Для обработки документов необходимо реализовать операцию выделения произвольной части документа (назовём её *операцией получения диапазона*), входным параметром которой является произвольное множество вершин графа, а результатом – подграф, порождённый этим множеством вершин. *Операция расшифровки* – «наложение» структуры на фрагмент (вершину графа). Помимо структуры и содержания, в большинстве приложений важную роль играют *визуальное оформление документа* и его представление в определенном формате, поэтому необходима и операция *представления документа в определенном формате*, представляющая функцию, задающую соответствие между фрагментами документа и некоторым множеством форматов, элементы которого задают правила отображения фрагментов. Операция *поиска* применима к различным составляющим документа: структуре, содержанию и представлению, а результатом операции будут фрагменты документа, удовлетворяющие заданным критериям поиска.

Многоаспектная онтология электронных документов. Для решения задач обработки электронных документов необходимо иметь консолидированные знания об их структуре и содержании (*формат* и *тип* электронного документа, его *структура* (состав)). При создании онтологического ресурса в него включаются понятия, относящиеся ко всем трём выделенным аспектам представления информации о документах, каждый из которых описывается онтологией, однако понятия, относящиеся к различным аспектам, связаны между собой. Таким образом, создаётся единая онтология электронных документов. Ресурс должен поддерживать возможность расширения и уточнения для настройки на решение задач, возникающих при обработке документов в различных ИС в течение всего их жизненного цикла.

Представление логики обработки документа. Задача добавления в электронные документы логики обработки данных решается на протяжении многих лет (реализуются как промышленные, так и исследовательские проекты). Среди промышленных решений можно выделить средства офисного программирования и языки макропрограммирования; добавление контекстно-зависимой логики в документы; решения для автоматизации заполнения электронных форм (Microsoft InfoPath, Google Forms).

Обработка электронных документов является одной из основных функций любой ИС, но в подавляющем большинстве случаев документам отводится пассивная роль: они являются «контейнером» для информации, предназначенной для пользователей. Однако, если добавить документам «активности», можно оптимизировать решение ряда традиционных для ИС задач. Задачу «активизации» документов предлагается решать в рамках *агентного подхода* [4, 5]. Для задания «динамики» документа предлагается использовать специализированные интеллектуальные агенты, представляющие «интересы» документа. Реализация подхода позволит не только снизить трудоёмкость задач, решаемых на различных этапах жизненного цикла ИС, но и предложить новые возможности для их разработчиков и пользователей.

Основная идея подхода заключается в том, что каждый документ в системе имеет связанного с ним *интеллектуального агента*, обладающего полной информацией о документе и способного «представлять интересы» документа при решении широкого круга задач. Для интеллектуального управления документами необходимо, чтобы агент, представляющий интересы документа, обладал сведениями о семантике документа, мог адекватно интерпретировать содержащуюся в документе информацию. Для этого необходимо решить задачу *семантического индексирования*. В [3] предлагается подход к решению этой задачи и описывается метод включения полученного семантического индекса непосредственно в тело документа. В рамках данного подхода предполагается, что основные понятия предметной области описаны в онтологии, и эта онтология доступна агентам и интерпретируема ими.

Поддержка жизненного цикла документа. На основе предлагаемого подхода планируется автоматизировать поддержку жизненного цикла (рис. 1) документов в ИС, а также снизить трудоёмкость анализа предметной области, происходящих в ней изменений при проектировании и сопровождении ИС:

- при анализе проектной документации могут быть выделены ключевые даты и этапы проекта и составлен план работ;
- на основе семантического индексирования документов, их интеллектуального анализа может быть выделен список сущностей предметной области и их атрибутный состав, ограничения, связи и выполняемые ими или над ними операции. Эта информация может быть использована системными аналитиками при разработке модели предметной области;
- информация об изменении содержания документов в процессе эксплуатации ИС может быть использована при обновлении модели предметной области информационной системы;
- информация об изменении шаблонов может быть использована при обновлении представления документов в системе;
- информация об изменении объектов информационной системы может быть использована для генерации и обновления содержания документов;
- информация об изменении реализуемой в системе бизнес-логики может быть использована для изменения содержания документов (например, инструкций и руководств).

Решение этих задач основано на общем подходе к активизации документов с использованием средств их интеллектуальной обработки.

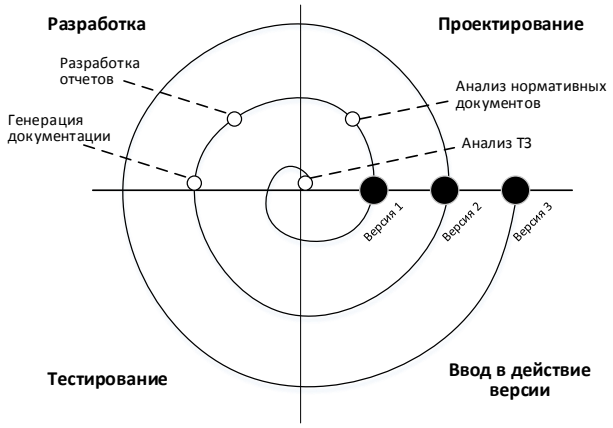

Рис. 1. Жизненный цикл ИС и документов

Заключение. Представленная модель позволяет формализовать алгоритмы интеллектуальной обработки электронных документов в информационных системах, предоставляя широкие возможности для интеграции документов с онтологическими ресурсами. Агентный подход позволяет решить проблему добавления бизнес-логики в документы. В отличие от других подходов в данном случае к документам не добавляется программный код и не расширяется их атрибутный состав, включаемая в документы информация. Предлагаемый подход позволяет абстрагироваться от технологических особенностей обработки и специфики конкретных форматов документов. Система становится более гибкой, интеллектуальной, адаптируемой к изменению среды. Данный подход позволяет решить широкий спектр задач, связанных с использованием электронных документов в ИС на всех этапах их жизненного цикла, что позволяет говорить о *документно-ориентированной парадигме* поддержки жизненного цикла ИС.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ланин В*. Онтологии как основа функционирования систем обработки электронных документов // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории». – Новосибирск, 2009. – Т. 2. – С. 173-177.
2. *Berners-Lee T., Hendler J., Lassila O*. The Semantic Web // Scientific American. – 2001. – Vol. 284 (5). – P. 34-43.
3. *Bessonov V., Lanin V, Sokolov G*. A semantic indexing of electronic documents in open formats // International Journal “Information Theories and Applications”. – 2012. – Vol. 19, № 2. – P. 139-148.
4. *Ginsburg M*. An Agent Framework for Intranet Document Management // Autonomous Agents and Multi-Agent Systems. – 1999. – Vol. 2, № 3. – P. 271-286.
5. *Pešović D., Vidaković M., Ivanović M., Budimac Z., Vidaković J*. Usage of agents in document management // ComSIS. – 2011. – Vol. 8, № 1. – P. 193-210.

Статью рекомендовал к опубликованию д.ф.-.м.н., профессор О.В. Русаков.

Ланин Вячеслав Владимирович – Национальный исследовательский университет «Высшая школа экономики»; e-mail: Lanin@Perm.ru; 614060, г. Пермь, бул. Гагарина, 37а; тел. +73422825372; кафедра информационных технологий в бизнесе; старший преподаватель.

Lanin Viacheslav Vladimirovich – National Research University Higher School of Economics; e-mail: Lanin@Perm.ru; 37a, Gagarina street, Perm, 614060, Russia; phone: +73422825372; the department of business information technologies; senior teacher.