УДК 519.689.3

ТЕХНОЛОГИЯ АНАЛИЗА ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ ПОДДЕРЖКИ НАУЧНОЙ И ПРОИЗВОДСТВЕННОЙ ДЕЯТЕЛЬНОСТИ*

Ю. А. Загорулько, Е. А. Сидорова

Институт систем информатики им. А. П. Ершова СО РАН,
630090, г. Новосибирск, просп. Академика Лаврентьева, 6
E-mail: zagor@iis.nsk.su

Рассматриваются особенности информационных систем поддержки научной и производственной деятельности, основанных на онтологиях. Обосновывается набор инструментальных средств, необходимых для поддержки анализа документов с целью автоматического извлечения из их текстов значимой для пользователя информации. Эти инструментальные средства должны поддерживать различные стадии анализа текста и включать рабочие места для создания специализированных словарей и баз знаний. Предлагаемый подход позволяет осуществлять семантический анализ текста на основе знаний о предметной области, жанре документа и особенностях терминологии, используемой специалистами данной области.

Ключевые слова: информационная система, онтология, семантический анализ текста, извлечение фактов, контент документа.

Введение. Большой объем накопленной информации и высокая скорость поступления новой предъявляют все более жесткие требования к современным информационным системам (ИС). Современная ИС должна не только обеспечивать хранение, тематическую рубрикацию документов и их поиск по ключевым словам, но и предоставлять пользователю непосредственный доступ к информации, содержащейся в документах. Для решения этой задачи необходим переход на качественно новый уровень обработки информации — семантический, который позволяет учитывать смысл (содержание) документов, извлекать из них важные для пользователя факты, а также обеспечивать их хранение и поиск.

За несколько последних лет это направление в информационных технологиях получило широкое распространение [1]. Созданные на основе семантических технологий ИС отличаются от традиционных тем, что используют знания о предметной области, выраженные в виде онтологий [2, 3], которые используются как для представления информации в виде взаимосвязанных фактов и решения задач информационного поиска [4], так и при анализе текста документов [5, 6]. В связи с этим особую актуальность приобретает разработка технологии анализа документов в контексте ее применения в информационных системах [7].

Важным аспектом, который нужно учитывать при разработке ИС, служащих для поддержки научной и производственной деятельности, является требование ее настраиваемости в процессе эксплуатации. Невыполнение этого требования может привести к тому, что система с течением времени перестанет выполнять свои функции из-за изменений в структуре предметной области или спектре требований, которые неизбежно ведут к изменению системы понятий, тематики документов и соответствующих условий их классификации и индексации.

*Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (проект № 09-07-00400).

Целью предлагаемой работы является представление подхода к анализу документов на основе знаний о предметной области и особенностях языка документов. Разработанная на основе данного подхода технология извлечения информации из текста ориентирована на документы деловой и научной тематики, что позволяет эффективно использовать ее в информационных системах поддержки научной и производственной деятельности.

Онтология в информационной системе. Основу системы знаний рассматриваемых ИС составляет онтология, формальное описание которой имеет вид

$$O = \langle C,\, R,\, T,\, D,\, A,\, F,\, Ax \rangle,$$

где $C = \{C_1, \ldots, C_n\}$ — конечное непустое множество классов, описывающих понятия некоторой предметной области; $R = \{R_1, \ldots, R_m\}$, $R_i \subseteq C \times C$, $R = R_T \cup R_P \cup R_A$, — конечное множество бинарных отношений, заданных на классах (понятиях): R_T — антисимметричное, транзитивное бинарное отношение наследования, задающее частичный порядок на множестве понятий C, R_P — транзитивное бинарное отношение включения («часть—целое»), R_A — конечное множество ассоциативных отношений; T — множество стандартных типов; $D = \{d_1, \ldots, d_n\}$ — множество доменов $d_i = \{s_1, \ldots, s_k\}$ (здесь s_i — значение стандартного строкового типа); $A = \{a_1, \ldots, a_w\}$ — конечное множество атрибутов, описывающих свойства понятий C и отношений R_A (конкретные значения атрибута a_i должны либо принадлежать одному из доменов $d_i \in D$, либо иметь тип $t_j \in T$; для каждого класса $C_i \in C$ экспертом выделяется подмножество ключевых атрибутов $A_i \subseteq A$, служащих для однозначной идентификации его объектов (экземпляров класса C_i)); F — множество ограничений на значения атрибутов понятий и отношений, т. е. предикатов вида $p_i(e_1, \ldots, e_m)$ (здесь e_k — либо имя атрибута ($e_k \in A$), либо константа ($e_k \in d_i$ или $e_k \in t_j$)); Ax — множество аксиом, позволяющих выводить дополнительные ассоциативные отношения между объектами:

$$\text{if}\, r_{p1}(c_1, c_2)\, \&\, \ldots\, \&\, r_{pn}(c_{m-1}, c_m)\, \text{then}\, r_c\,(c_k, c_l),$$

где $r_{pi} \in R,\ i \in 1, \ldots, n;\ c_j \in C_j \subseteq C, j \in 1, \ldots, m;\ r_c \in R_A;\ c_k, c_l \in C_j,\ k, l \in 1, \ldots, m.$

Онтология для конкретной ИС строится на основе введенного выше формализма согласно методологии, предложенной в [8]. Главным принципом этой методологии является построение требуемой онтологии на основе базовых онтологий путем их достройки и развития, что значительно упрощает создание онтологии для конкретной ИС и ее дальнейшее сопровождение. В качестве базовых выбраны три онтологии: онтология деятельности, которая составляет базис онтологии проблемной области ИС, онтология предметного знания, на основе которой строится онтология предметной области ИС, и онтология базовых задач ИС, используемая для построения онтологии задач ИС.

В современной информационной системе онтология выполняет следующие функции:

1. Описание предметной области информационной системы. Онтология представляет моделируемую предметную область в виде множества понятий и множества заданных на них отношений.

2. Спецификация структуры информационного наполнения (контента) ИС. Онтология, вводя формальные описания понятий предметной области ИС в виде классов объектов и отношений между ними, задает структуры для представления реальных данных и связей между ними. Организации, персоны, текстовые и мультимедийные документы, как и другие объекты, которые необходимо представить в базе данных (БД) информационной системы, могут быть описаны с помощью онтологии. Для этого в нее вводится понятие, характеризующее соответствующий тип объекта, определяется его структура и возможные взаимосвязи (отношения) с объектами других понятий.

3. Представление содержания (контента) документов [9]. На основе онтологии строится содержательная аннотация документа, включающая извлеченные из его текста объекты и связи, соответствующие понятиям и отношениям онтологии. Отметим, что с помощью онтологии описывается не все содержание документа, а лишь те его аспекты, которые существенны для решения конкретных задач в рамках данной системы.

4. Семантический поиск. Онтология обеспечивает поиск информации в терминах предметной и проблемной областей, т. е. пользователь может формулировать поисковые запросы, основными элементами которых являются понятия и отношения онтологии, а также ограничения, которым должны удовлетворять искомые данные.

5. Интеллектуальная интеграция информации. На основе онтологии может интегрироваться информация из различных информационных источников благодаря тому, что их содержание единообразно отображается в понятия и отношения общей для них онтологии.

Таким образом, использование онтологии в качестве основы ИС делает систему знаний легко расширяемой и настраиваемой — в нее могут интегрироваться как новые знания (например, о новых понятиях и отношениях предметной области), так и новые типы информационных ресурсов (например, новые типы документов).

Извлечение информации из текста. До недавнего времени задача анализа текста на естественном языке рассматривалась многими исследователями независимо от той обстановки, где планировалось использовать ее результаты. В отличие от работ, связанных с задачей полного извлечения смысла или извлечения всей информации из текстов документа, для большинства ИС нет необходимости делать полный семантический анализ всего связанного текста.

Важным с точки зрения анализа документа свойством онтологии является то, что она задает формат хранения данных в системе, а следовательно, определяет информацию, которую необходимо извлекать из текста документа или игнорировать. Результат анализа документа представляется в виде семантической сети объектов, которые являются экземплярами понятий и отношений онтологии предметной области.

Очевидно, что знаний о предметной области, имеющихся в онтологии, недостаточно для автоматического извлечения информации из текста, поэтому требуются дополнительные знания о языке, на котором эта информация представлена, и, следовательно, программные компоненты, обеспечивающие формирование таких знаний экспертами (лингвистами) и автоматическое применение полученных знаний при обработке документов.

Архитектура системы анализа документов. Система анализа документов включает пять независимых компонентов (см. рисунок):

1) словарный компонент, включающий автоматизированное рабочее место (АРМ) настройки словарей разных типов и исполняемые модули, реализующие основные методы словарной обработки текста;

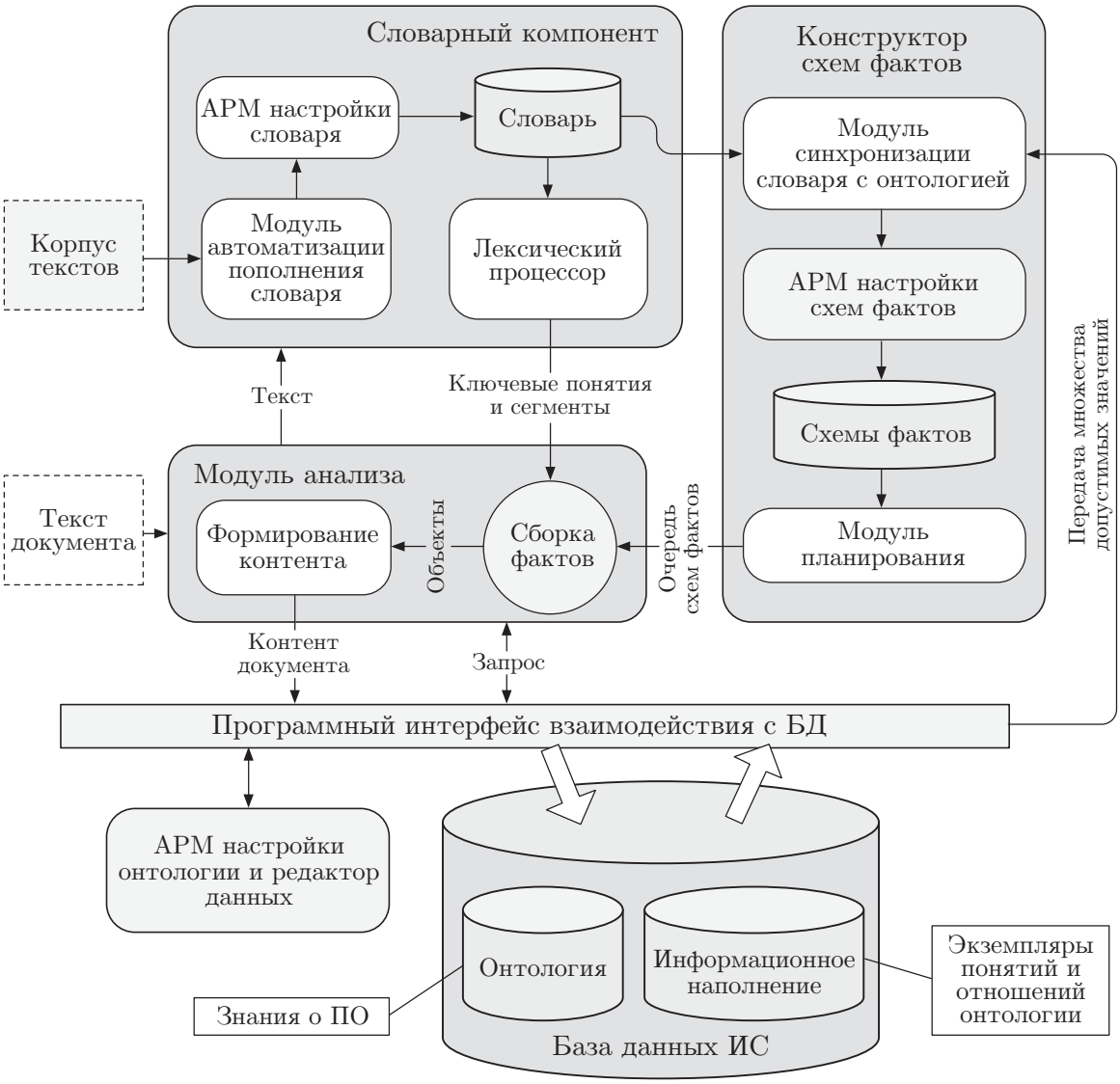
2) конструктор схем фактов, обеспечивающий настройку процесса постсловарного анализа документа;

3) модуль анализа — исполняемое ядро, осуществляющее анализ текста документа, начиная со словарного анализа (с помощью словарного компонента), сегментации, извлечения фактов по заданным схемам и заканчивая процессом формирования контента документа;

4) АРМ настройки онтологии, обеспечивающее настройку контента ИС (в частности, определяется формат извлекаемой из документов информации);

5) программный интерфейс взаимодействия с базой данных ИС, предоставляющий унифицированный доступ к знаниям (онтологии) и данным (контенту).

Система работает в двух режимах: настройки базы знаний и обработки документов.



Архитектура системы анализа

Настройка базы знаний. В режиме настройки эксперт с помощью соответствующих АРМ формирует базу знаний ИС. При этом ключевая лексика для словарей и онтологий автоматически извлекается из подборки документов, релевантных предметной области ИС.

В ходе построения базы знаний необходимо решать задачу согласования лингвистических знаний, заданных в словаре, со знаниями, представленными онтологией ИС. При этом нужно иметь в виду, что содержащееся в тексте документа описание объекта, соответствующее тому или иному элементу онтологии, редко однозначно представляется с помощью одного термина, устойчивого словосочетания или другой единицы словаря. Такое описание обычно включает несколько компонентов, распределенных по тексту. Для того чтобы понять, о каком элементе онтологии идет речь в тексте, требуется выделить факты, явным образом определяющие класс и/или атрибут понятия или отношения.

Предлагается использовать декларативное описание значимого для ИС факта, связывающее элементы словаря с понятиями и отношениями онтологии. Такую структуру будем называть схемой факта.

Схема факта F_k — это тройка вида $\langle \mathrm{Arg}, Rs, C \rangle$, где Arg — множество дескрипто-

ров аргументов факта (здесь дескриптором может быть тип словарной единицы, класс информационного объекта (понятие или отношение онтологии) или тип факта, играющего вспомогательную роль при описании цепочки, связывающей термины с понятиями и отношениями онтологии); $Rs = \langle t, \mathrm{op}(t), P\rangle$ — результат применения схемы (здесь t задает тип элемента (класс нового объекта или один из аргументов), $\mathrm{op}(t)$ — тип операции (создание и/или редактирование аргумента), применяемой при условии выполнения ограничения C, P — множество правил для формирования/редактирования объекта, каждое правило назначает атрибуту результирующего объекта либо конкретное значение соответствующего типа данных, либо значение атрибута одного из аргументов факта); C — множество ограничений, накладываемых на характеристики аргументов факта (условия на атрибуты аргументов, структурно-текстовые и синтаксические ограничения).

Схемы фактов создаются с помощью конструктора, который предоставляет доступ к словарям и онтологии и обеспечивает корректность построенных схем. Модуль планирования формирует из схем очередь, определяющую порядок применения схем в процессе анализа с минимальной потерей информации.

Автоматическая обработка документа. В режиме обработки на вход модуля анализа поступает текст документа, который передается лексическому процессору словарного компонента. Словарный компонент осуществляет сегментацию и извлечение ключевых терминов и передает результаты своей работы модулю сборки фактов, который осуществляет поиск фактов в тексте, используя упорядоченный набор схем фактов. Результатом работы модуля сборки фактов является множество информационных объектов. Модуль формирования контента документов идентифицирует и уточняет параметры полученных объектов, сравнивая их с информационными объектами, хранящимися в БД ИС, формирует информационные объекты, представляющие в БД проанализированный документ и его контент, а также устанавливает между ними необходимые связи (экземпляры отношений).

Рассмотрим подробнее последний компонент, поскольку именно он обеспечивает взаимодействие подсистемы анализа с контекстом самой ИС и позволяет существенно расширить возможности анализа, связанные с уточнением и корректировкой полученной из документа информации.

Для того чтобы сформировать контент документа, необходимо:

— обеспечить корректность объектов, полученных в результате анализа;

— идентифицировать объекты, образующие контент;

— добавить объекты в информационное пространство системы и связать их с документом.

Объект считается корректным, если определен его класс и значения всех установленных (извлеченных из текста) атрибутов удовлетворяют заданным в онтологии ограничениям (в частности, принадлежат соответствующим доменам или типам).

Объект считается идентифицированным, если определены его класс и значения всех ключевых атрибутов. Данное свойство позволяет однозначно выделить этот объект из множества других объектов, т. е. обеспечивает его уникальность в БД системы.

Обеспечение корректности объектов. Корректность значений атрибутов объектов обеспечивается на уровне словаря и схем фактов. На этапе конструирования словаря эксперт должен либо внести в него термины, обозначающие доменные значения, либо предусмотреть в схемах фактов возможность создания объектов, атрибуты которых позволят воссоздать название доменного значения. На этапе анализа в результате применения соответствующей схемы факта термины преобразуются в строковые (доменные) значения атрибутов.

Поскольку описание одного и того же доменного значения в тексте может осуществляться с помощью различных терминов, то необходимо предусмотреть механизм сопостав-

ления разных терминов с одним значением. Для этого в словарной подсистеме определяется синонимичная группа — набор терминов, альтернативных данному значению.

Идентификация объектов. В процессе идентификации осуществляется уточнение полученных при анализе объектов (уточнение атрибутов объектов и их связей), «склеивание» одинаковых объектов на основе использования локального и глобального контекстов, а также поиск объектов в информационном пространстве системы.

Под локальным контекстом понимается содержание анализируемого документа, которое используется для решения следующих задач:

— определения референта для местоимений;

— определения референта для неоднозначных именных групп;

— отождествления объектов, имеющих один и тот же референт (который обозначает один и тот же объект действительности), на основании частично совпадающего и непротиворечивого набора атрибутов и связей.

В данном случае уместно говорить о задаче разрешения анафоры, которая стала весьма популярной в области компьютерной обработки текста. Например, алгоритмы автоматического разрешения анафоры описываются такими авторами, как Г. Хирст, Ш. Лаппин, Р. Митков, М. Поэсио и др. (см., например, [10]). Современные системы разрешения анафоры работают с эффективностью от 60 до 90 %.

Глобальный контекст представлен всем информационным пространством системы, в котором осуществляется поиск объекта. Он содержит как результаты ранее проанализированных документов, так и информацию, вводимую пользователем вручную или поступающую из подсоединенных баз данных.

Использование глобального контекста возможно в случае, когда набор ключевых атрибутов задан не полностью. В этом случае осуществляется поиск максимально похожего объекта в БД системы по известному набору атрибутов и классу объекта, а также его связям.

Предложенный способ заключается в построении фокусного множества объекта, найденного в тексте, которое включает все объекты, непосредственно связанные с данным с помощью экземпляров отношений, и сопоставлении его с фокусными множествами объектов, найденных в БД системы. Отметим, что в фокусное множество имеет смысл включать уже идентифицированные в БД объекты.

Отметим еще две возможности использования глобального контекста.

1. Уточнение класса объекта по иерархии классов (при этом может уточняться как объект, найденный в тексте, так и объект из БД системы).

2. Идентификация и уточнение объектов с помощью восстановления иерархии вложенности объектов по отношению часть—целое.

Качество работы системы анализа. Качество работы системы анализа оценивается как расхождение автоматически сгенерированного контента документа с контентом, построенным экспертом. Оценка осуществляется на представительном массиве документов. Для этого предлагается применять принятые в области автоматической обработки текстов показатели полноты и точности [11].

Полнота (P) — отношение количества правильно идентифицированных системой объектов, образующих контент документа, к числу объектов, определенных экспертом.

Точность (T) — отношение количества правильно идентифицированных объектов, образующих контент документа, к общему числу объектов, определенных системой.

Для оценки качества необходимо дополнительно учитывать:

— неключевые атрибуты объекта (их правильность/ошибочность);

— неточно или частично определенные объекты (например, класс объекта не доуточнен или не определен один из ключевых атрибутов, для которого допустимо пустое значение);

— ошибочно идентифицированные объекты (например, объект неправильно сопоставлен с объектом БД или есть ошибка в значении ключевого атрибута).

Будем рассматривать экземпляр отношения как объект, к ключевым атрибутам которого относятся аргументы отношения.

Показатели полноты и точности запишем как

$$P = \frac{K1 + \beta \sum\limits_{i=1}^{K1} \frac{k_i}{n_i} + \gamma\Big(K2 + \beta \sum\limits_{j=1}^{K2} \frac{k_j}{n_j}\Big)}{(1+\beta)(K1 + \mathrm{Miss} + K2)}, \quad T = \frac{K1 + \beta \sum\limits_{i=1}^{K1} \frac{k_i}{m_i} + \gamma\Big(K2 + \beta \sum\limits_{j=1}^{K2} \frac{k_j}{m_j}\Big)}{(1+\beta)(K1 + \mathrm{Error} + K2)},$$

где $K1$ — число правильно идентифицированных объектов; $K2$ — число частично идентифицированных объектов; Miss — число пропущенных объектов; Error — число ошибочно идентифицированных объектов; k_i — число правильно определенных системой неключевых атрибутов i-го объекта; m_i — число всех неключевых атрибутов i-го объекта, определенных системой; n_i — число неключевых атрибутов i-го объекта, определенных экспертом; β $(0 \le \beta \le 1)$ — коэффициент важности неключевых атрибутов; γ $(0 \le \gamma < 1)$ — коэффициент значимости частично определенных объектов.

Заметим, что приведенные выше формулы оценки полноты и точности не учитывают различия в весах атрибутов, а также различную значимость объектов в зависимости от их типов.

Заключение. Описанная в данной работе технология анализа документов предоставляет эргономичные средства создания онтологий, предметных словарей и схем фактов, с помощью которых настройка процесса содержательной обработки документов может выполняться непосредственными носителями знаний — экспертами и лингвистами, не имеющими специальных навыков программирования.

Основные компоненты предлагаемой технологии были успешно апробированы в ряде практических приложений, служащих для поддержки научной и производственной деятельности, в частности при разработке интеллектуальной системы документооборота инвестиционной компании [12] и портала знаний по археологии [13].

СПИСОК ЛИТЕРАТУРЫ

1. **Хорошевский В. Ф.** Управление знаниями и обработка ЕЯ-текстов // Тр. IX Нац. конф. по искусственному интеллекту (КИИ-2004). М.: Физматлит, 2004. Т. 2. С. 565–572.
2. **Gruber T. R.** Toward principles for the design of ontologies used for knowledge sharing // Intern. Journ. Hum.-Comput. Studies. 1995. **43**, Is. 5–6. P. 907–928.
3. **Guarino N.** Formal ontology in information systems // Proc. of the FOIS'98. Amsterdam: IOS Press, 1998. P. 3–15.
4. **Загорулько Ю. А., Боровикова О. И.** Подход к построению порталов научных знаний // Автометрия. 2008. **44**, № 1. С. 100–110.
5. **Хорошевский В. Ф.** OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов // Тр. IX Нац. конф. по искусственному интеллекту (КИИ-2004). М.: Физматлит, 2004. Т. 2. С. 573–581.
6. **Нариньяни А. С.** NLP: технологическая база // Тр. XI Нац. конф. по искусственному интеллекту с международным участием (КИИ-2008). М.: ЛЕНАНД, 2008. Т. 3. С. 225–233.
7. **Рубашкин В. Ш.** Семантический компонент в системах понимания текста // Тр. X Нац. конф. по искусственному интеллекту с международным участием (КИИ-2006). М.: Физматлит, 2006. Т. 2. С. 455–463.

8. **Загорулько Ю. А., Боровикова О. И.** Технология построения онтологий для порталов научных знаний // Вест. НГУ. Сер. Информационные технологии. 2007. **5**, вып. 2. С. 42–52.
9. **Васильев И. А., Тузовский А. Ф.** Структура системы управления знаниями // Тр. Междунар. симп. «Информационные и системные технологии в индустрии, образовании и науке». Караганда: Издательство КарГТУ, 2003. С. 286–288.
10. **Mitkov R.** Anaphora resolution // The Oxford handbook of computational linguistics /Ed. R. Mitkov. N.Y.: Oxford university press, 2003. P. 266–283.
11. **Хорошевский В. Ф.** Оценка систем извлечения информации из текстов на естественном языке: кто виноват, что делать // Тр. X Нац. конф. по искусственному интеллекту с международным участием (КИИ-2006). М.: Физматлит, 2006. Т. 2. С. 464–478.
12. **Загорулько Ю. А., Кононенко И. С., Сидорова Е. А., Костов Ю. В.** Подход к интеллектуализации документооборота // Информационные технологии. 2004. № 11. С. 2–11.
13. **Андреева О. А., Боровикова О. И., Загорулько Ю. А. и др.** Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Тр. X Нац. конф. по искусственному интеллекту с международным участием (КИИ-2006). М.: Физматлит, 2006. Т. 3. С. 832–840.

Поступила в редакцию 2 апреля 2009 г.