УДК 519.766.2

Н. Ф. ХАЙРОВА, канд. техн. наук, доц., НТУ «ХПИ»;
В. А. ТАРЛОВСКИЙ, асп. ХНТУ, г. Херсон

ИСПОЛЬЗОВАНИЕ СЕМАНТИКО-ОРИЕНТИРОВАННОГО ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ДЛЯ ДОБЫВАНИЯ НОВЫХ ЗНАНИЙ ИЗ ПОТОКА ДОКУМЕНТОВ КОРПОРАТИВНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

У статті пропонується модель перекладу різноформатної текстової інформації в інтелектуальні активи компанії. Розглядається алгоритм роботи семантико-орієнтованого лінгвістичного процесора, що отримує знання з документів, які надходять до організації. Послідовно описані предлінгвістічний, графемний, морфологічний, контекстний, статистичний і логіко-алгебраїчний етапи роботи процесора.

В статье рассматривается модель перевода разноформатной текстовой информации в интеллектуальные активы компании. Предлагается алгоритм работы семантико-ориентированного лингвистического процессора, извлекающего знания из поступающих в организацию документов. Последовательно описаны предлингвистический, графемный, морфологический, контекстный, статистический и логико-алгебраический этап работы процессора.

In the article there has been proposed model of transfer from textual information in various formats into the company's intellectual assets. In this work there has been showed the algorithm of work of the semantic linguistic processor extracting knowledge from documents of the organization. The semantic linguistic processor has to contain a predlingvisticheskу stage, a graphemic stage, a morphological stage, a context-dependent stage, a statistical stage and algebraic logic stage.

Введение. Рассматривая современные корпоративные информационные системы (КИС) с точки зрения управление знаниями можно заметить, что сегодня основной акцент делает не на сохранение разрозненной информации, а на извлечение закономерностей и принципов, позволяющих решать производственные и бизнес задачи, т.е. осуществлять накопление знаний [1, 2]. Источником знаний являются различные документы, поступающие в систему на обработку — это корпоративные стандарты, методики, бизнес-правила и технологии, Технологическая и трудовая документация, накопившаяся в процессе функционирования предприятия. При чем, перед КИС ставится задача извлечь и накопить именно инновационные знания, которые снабжают фирму конкурентным потенциалом, а не коренные, устоявшиеся и даже «старые» знания, которые имеют все участники данной отрасли.

Таким образом, основной задачей повышения эффективности КИС становится разработка системы трансформации информации, доступной снаружи и внутри организации в интеллектуальные архивы компании, представляющие инновационные знания. И если для решения данной задачи при работе со структурированной информацией используются хорошо разработанные технологии «добычи данных» (data mining), то для обработки слабоструктурированной и неструктурированной информации необходимо разрабатывать единое информационное пространство, представляющее модель знаний предметной области работы КИС.

Постановка задачи. Необходимо разработать подсистему КИС, решающую задачи логического анализа (e-analytics) неструктурированной информации для построения базы знаний, которая должна наряду с архивами и прочими полнотекстовыми массивами, прорабатывать поступающие в корпорацию новые электронные документы. Подсистема должна учитывать, что поступающая в организацию информация (электронная почта, мгновенные сообщения, HTML-документы, XML-документы, электронные документы других форматов) обычно находится в беспорядочном состоянии и представлена в документах различных типов и форматов.

Описание модели. Предлагаемая модель перевода разнородной электронной информации в интеллектуальный актив компании, представляет собой семантико-ориентированный лингвистический процессор, включающий семь этапов:

- предлингвистическая обработка;
- графемная обработка;
- морфологическая обработка;
- контекстный анализ;
- статистическая обработка;
- логико-алгебраическая обработка обучающей выборки;
- динамическая классификация поступающей информации.

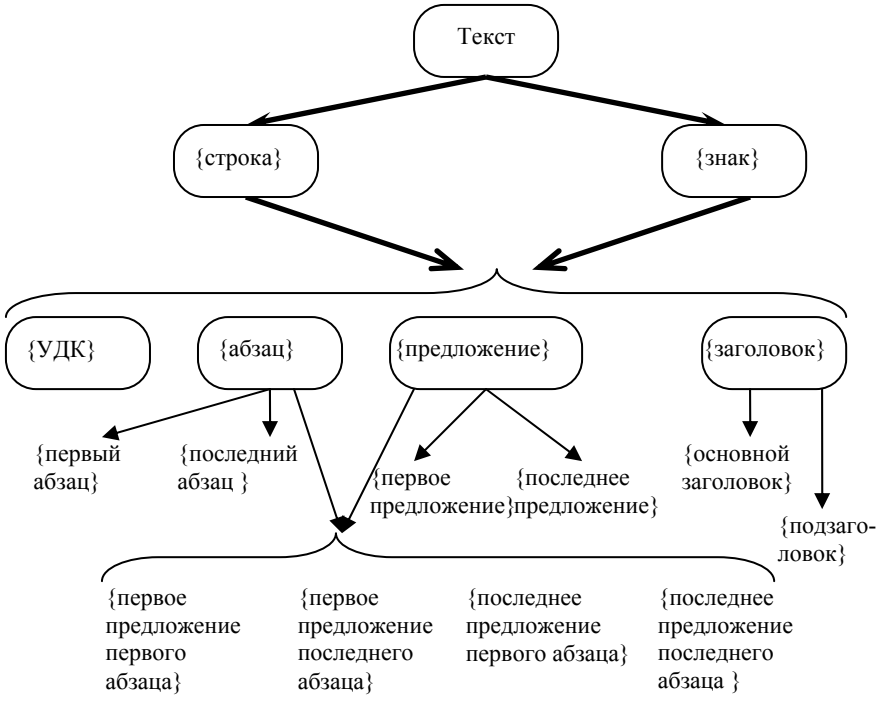
Предлингвистическая обработка. На вход предлингвистического этапа анализа поступают документы широкого спектра форматов. На данном этапе выделяются документы, имеющие в полях иерархических спецификаций списки ключевых слов и словосочетаний документа.

В HTML-документах удаляются описания стилей, сценарии и т.д. Удаляются все HTML-тэги, кроме тегов определяющих структурное деление документов. С помощью мета определителей <META NAME="subject"> <META NAME= "keywords"> и других выделяются ключевые слова и тема документа, сформулированные автором веб-страницы.

Графемная обработка. На вход графемного анализа поступает множество текстов $T=\{t_1, t_2, \ldots, t_n\}$ документов в виде полнотекстовой базы данных. Текст на естественном языке можно представить как целостный объект, элементами, которого являются знаки, организованные определенным образом в строке [3]: ТЕКСТ = {{знак}, {строка}}.

Задача графемного этапа обработки состоит в выделении единиц текста, имеющих графемное значение и выделение парадигматических отношений

данных единиц, позволяющих соотнести их с некоторым классом, который отражает одинаковые свойства выделенных элементов (см. рисунок).



Логико-структурная модель графемного значения текста

Морфологическая обработка. Блок морфологической обработки вводится для учета словоизменительных форм и представления слова в канонической форме. Обычно, ключевыми словами (КС) документа являются понятия об объектах информации, обозначаемые существительными (иногда вместе с определениями), в канонической форме. Определением в текстах на украинском и русском языках выступают прилагательные, причастия, порядковые числительные или существительные в родительном, творительном, дательном и предложном падежах.

Морфологическая обработка осуществляется методом квазиокончаний. На вход поступает множество L всех лексем текста в соответствующей форме, за исключением слов отрицательного словаря. Конечный сегмент каждой словоформы из множества L проверяется на совпадение со словарем квазиокончаний, начиная от трех конечных буквосочетаний. В случае их совпадения квазиокончание у словоформы отсекается, и основа помещается в словарь квазиоснов $L' = \{l'_1, l'_2, ..., l'_k\}$. Словарь окончаний представляет собой множество окончаний прилагательных, причастий, порядковых числительных единственного и множественного числа, существительных в родительном, творительном, дательном и предложном падежах соответствующего языка (русского или украинского).

Для аналитического английского языка (языка со слаборазвитой морфологией) у каждой словоформы из множества L, отсекается конечный сегмент, в случае его принадлежности множеству квазиокончаний {-a, -e, -ed, -er, -es, -est, -en, -iest, -ing, -is, -on, -s, -s', -'s, -t, -m, -y, -ying, -ies}. Правая, оставшаяся часть словоформы помещается в словарь квазиоснов. В случае отсутствия квазиокончания ключевое слово l_i переносится в словарь L' целиком $L' = \{l'_1, l'_2, ..., l'_k\}$ $(l_i = l'_i,\ i \in \{1, 2, ..., k\})$.

На этапе контекстного анализа из множества лексем текста L определяются словосочетания. Словосочетаниями считаются два или больше последовательно расположенных слов текста, синтаксически связанных по типу управления или согласования. Для определения словосочетаний выбираются рядом стоящие существительное и определение, обладающие эквивалентной морфологической информацией или несколько рядом стоящих существительных, все кроме одного из которых имеют грамматическую форму родительного или творительного падежа.

На следующем *статистическом этапе* работы алгоритма система получаем информационное представление каждого документа в виде множества ключевых слов и словосочетаний текста $L_m = \{t_m^i\}, 1 \le i \le n$, UDK-текста u_m и множества значений рубрикатора данного текста, $R_m = \{r_m^i\}, 1 \le i \le n$, где m — номер документа, поступающего в корпоративную информационную систему. Информационное представление текста — это структура, с одной стороны, отображающая суть документа, назначение и взаимосвязь его составляющих, и, с другой стороны, показывающая отношение документа, поступающего на обработку в КИС предметной области деятельности менеджера [4].

Для определения алфавитного словаря ключевых слов документа, не содержащего соответствующие поля мета информации, на этапе статистической обработки формируется множество квазиоснов L'_m, каждая из которых проверяется на совпадение слева со всеми словоформами текста.

Для учета информационной значимости ключевых слов вводим весовые коэффициенты, являющиеся дополнительным средством семантической дифференциации лексических единиц документа (ЛЕД). Алгоритм определения веса ключевого слова l'_i базируется на гипотезе зависимости информационной значимости ЛЕД от ее "позиции" в тексте, т.е. ее принадлежности к тем или иным структурно-определенным, найденным на

этапе графемной обработки, фрагментам текста. Используем весовые коэффициенты (v), предложенные в работе [5].

Если на этапе контекстного анализа в документе обнаружены словосочетания то типу согласование $A+N$ или управление, $N = N_{род}, N + N_{тв}$, то их вес вычисляется как коэффициент, менее редкого слова входящего в словосочетание [6]. Далее словосочетания рассматриваются наряду с другими ключевыми словами.

В результате проверки всех словоформ текста документа d_j на совпадение с множеством квазиоснов L', получаем заполненную базу данных словаря ключевых слов документа d_j, в которой каждой встретившейся в тексте основе из словаря основ ключевых слов L' приписан вес данной основы в тексте. Формируем множество ключевых слов и словосочетаний документа, включающее основы слов и словосочетаний, имеющие наибольший вес.

Таким образом, в $L_m = \{l_m^i\}$ каждого текста помещают набор морфологически нормализованных наиболее информативных слов и словосочетаний, отобранных из текста, отражающих его основное предметное содержание.

Логико-алгебраическая обработка обучающей выборки. Используя полученное на предыдущих этапах работы системы информационное представление каждого документа в виде множества ключевых слов и словосочетаний текста $L_m = \{l_m^i\}, 1 \le i \le n$, UDK-текста u_m, значение рубрик данного текста r_m^i, описываем связь предметной области деятельности менеджера, работающего с документами, и предметных переменных, объективно определяющих глубинные знания документа. Для этого строим бинарные предикаты: $P_l(l,m), P_r(r,m), P_u(u,m)$ [5].

Предикат $P_l(l,m)$, заданный на декартовом произведении $M \bullet L$, характеризует отношения между множеством областей деятельности менеджеров организации и ключевыми словами рассматриваемых документов. Предикат $P_l(l,m) = 1$, когда документ, содержащий ключевое слово l, относится к области деятельности менеджера m.

Предикат $P_r(r,m)$, заданный на декартовом произведении $M \bullet R$, характеризует отношения между множеством областей деятельности менеджеров организации и значениями предметных рубрик рассматриваемых документов. Причем $P_r(r,m) = 1$ тогда и только тогда, когда документ, относящийся к предметной рубрике r, относится к области деятельности менеджера m.

Предикат, заданный на декартовом произведении $M \bullet U$, характеризует отношения между множеством областей деятельности менеджеров организации и значениями УДК. Причем, $P_u(u,m) = 1$ тогда и только тогда, когда документ, относящийся к УДК u, относится к области деятельности менеджера m и $P_u(u,m) = 0$, если документ, имеющий значение УДК — u, не относится к области деятельности менеджера m.

Предикаты P_l, P_r и P_u можно представить в виде таблиц, в ячейках которой ставятся единицы или нули в зависимости от того, равен ли соответствующий предикат единице или нулю для данных значений предметных переменных l, u, r, m.

На практике часто встречаются ситуации, когда, исключая из рассмотрения некоторые элементы декартового произведения, мы получаем разбиения множеств, более соответствующие интуитивным представлениям специалистов о семантике ключевых слов, предметных рубрик и значении УДК. Исключение некоторых элементов декартова произведения модели имеет смысл в том случае, когда упорядоченных пар немного по сравнению с общим числом элементов декартового произведения. Разработанный метод построения разбиений множеств L, R и U учитывает такие исключения. Алгоритм допускает небольшие различия между строками таблиц, попадающими в один класс. Мера таких допустимых отклонений ρ' может устанавливаться пользователем. При этом те строки (или столбцы), которые могут быть отнесены к различным классам, выделяются особо и могут быть классифицированы пользователем отдельно. Такие строки (столбцы) отличаются от элементов некоторых классов в числе ρ' двоичных разрядов, не превышающем заданного $\rho'(\rho \le \rho'$, где ρ можно интерпретировать как расстояние между векторами: $\rho'(a,b)\sum \alpha_i \oplus \beta_i$ a и b — двоичные векторы: $a = (\alpha_1, \alpha_2, ..., \alpha_k)$, $b = (\beta_1, \beta_2, ..., \beta_k)$

Система проводит анализ зависимости числа элементов разбиения от допустимого числа двоичных разрядов, в которых могут различаться элементы одного класса. Эта зависимость может использоваться для построения оптимальных разбиений, содержащих заданное число классов.

Разбивая, таким образом, ключевые слова обрабатываемых КИС документов, мы иерархически упорядочиваем их, используя для них отношение «быть элементом класса», включаться в предметную область исследования менеджера и указывая на их отношения к предметной рубрике и номеру УДК. Любые другие более сложные отношения на множестве, извлеченных из обрабатываемых документов ключевых слов, могут добавляться в базу знаний менеджером интеллектуально. При этом система может гибко и динамично менять разбиение ключевых слов, т.е. понятий и объектов, обрабатываемых менеджером в процессе решения управленческих задач.

Динамическая классификация поступающей информации. На этом этапе для каждого множества текстов документов, поступающих на вход системы, выполняются все описанные процедуры: предлингвистической, графемной,

морфологической, контекстной и статистической обработки. В результате каждому документу приписывается его информационное представление, которое сравнивается с имеющимися эталонами информационного представления, соответствующими области деятельности того или иного менеджера.

Выводы. Использование разработанного семантико-ориентированного лингвистического процессора позволяет извлекать из неструктурированных потоков текста на естественных языках семантическую информацию, представляемую множеством сведений о различных сторонах и отношениях изучаемых объектов заданной ПО, т.е. извлекать знания, динамически наполняя базу знаний информационной системы.

Рассматриваемая логико-лингвистическая модель одновременно с динамическим наполнением базы знаний новыми для данной предметной области понятиями и отношениями между этими понятиями, структурирует полнотекстовые информационные потоки, разбивая множество документов различных форматов, поступающее на обработку в КИС, на персонифицированные области деятельности каждого менеджера.

Список литературы: 1. *Гаврилова Т. А.* Базы знаний интеллектуальных систем. – СПб : Питер, 2000. – 384 с. **2.** *Mancini J.* Enterprise Content Management: Critical Technologies for Business Applications // AIIM, 2001. – P. 34–76. **3.** *Шаронова Н. В., Тарловский В. А., Хайрова Н. Ф.* Модель извлечения глубинных знаний для систем организационного управления. // Вестник Херсонского национального университета. – 2010. № 2 (38). С. 97–102. **4.** *Braslavski P., Shishkin A. A* User-Center Comparison of Web Search Engines. In Computational Linguistics and Intelligent Technologies. Proceedings of the Dialogua'2005 conference. Zvenigorod, June 1 – 6, 2005. P. 554–560. **5.** *Manning C., Schütze H.* Foundations of Statistical Natural Language Processing. MIT Press, 2000. – P. 28–57. **6.** *Хайрова Н., Шаронова Н.* Построение логической сети предиката персонификации области знаний менеджера. // Системный анализ и информационные технологии: материалы 12-й Международной научно–технической конференции SAIT 2010, Киев, 25–29 мая 2010 г. / УНК «ИПСА» НТУУ «КПИ». – К. : УНК «ИПСА» НТУУ «КПИ». 2010. – С. 333.

Надійшла до редколегії 06.11.2010