УДК 631.3

А.Н. СЕРЕБРОВСКИЙ

ОБ АВТОМАТИЗАЦИИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Abstract. *An approach to creating of knowledge extracting technology from text information resources is offered. The approach is based on the subject domain ontology and semantic marking out of initial data with the use of metadata.*
Key words: *ontology, semantic marking out, extraction of knowledge, information resource, thesaurus.*

Анотація. *Пропонується підхід до створення технології здобуття знань з текстових інформаційних ресурсів. Підхід ґрунтується на онтології предметної області і семантичній розмітці вихідних текстів з використанням метаданих.*
Ключові слова: *онтологія, семантична розмітка, здобуття знань, інформаційний ресурс, тезаурус.*

Аннотация. *Предлагается подход к созданию технологии извлечения знаний из текстовых информационных ресурсов. Подход основывается на онтологии предметной области и семантической разметке исходных текстов с использованием метаданных.*
Ключевые слова: *онтология, семантическая разметка, извлечение знаний, информационный ресурс, тезаурус.*

1. Введение

Приобретение знаний о предметной области (ПрО) является одной из необходимых функций при формировании, пополнении и актуализации базы знаний (БЗ) экспертных систем (ЭС). Извлечение знаний о ПрО из информационных ресурсов (ИР) – достаточно трудоемкий процесс, поэтому совершенствование системных средств, направленных на автоматизацию подобных процедур, является актуальным для решения проблем создания и функционирования ЭС. Одним из направлений решения задачи извлечения знаний является предварительная ручная обработка информационных ресурсов, в результате которой создается формализованное описание их семантики. Однако реальные объемы ИР таковы, что значительная их часть (в том числе WEB ИР) не обеспечивается семантическим описанием. Не решает также проблему целевого извлечения знаний в среде WEB технология поиска ИР по ключевым словам. “Поиск в WEB традиционными способами (например, по ключевым словам) стал малоэффективным” [1]. Один из системных подходов решения проблемы снижения трудоемкости извлечения знаний из текстовых ИР предлагается в рамках направления Semantic Wiki [2]. В результате применения данного подхода формируются тексты, которые, наряду с исходной информацией, содержат данные о знаниях, излагаемых в ИР (понятиях ПрО и их отношениях). При этом степень подробности описания знаний определяется:

– предварительно построенной конкретной онтологией ПрО;

– системой запросов, с помощью которых пользователь извлекает нужные знания из текстовых ИР.

С помощью данной технологии можно описывать не только полный текст ИР, но и любые его фрагменты, например, каждую страницу. В этом случае весь текст ИР становится размеченным на фрагменты, каждый из которых дополняется специфическими для него метаданными.

В настоящей работе предлагается подход к созданию одного из вариантов технологии извлечения знаний в рамках Semantic Wiki. Подход основан на предварительной семантической разметке исходных текстов понятиями конкретной онтологии узкой предметной области “Оценка и

ISSN 1028-9763. Математичні машини і системи, 2010, № 3

анализ пожаровзрывоопасности ситуаций на объектах типа “Хранилища взрывоопасных предметов (ХВП)”.

2. Онтология ПрО

Онтология – результат детальной формализации конкретной области знаний, в рамках которой выполняется поиск [3].

Онтология включает в себя:

– понятия проблемной области знаний;

– связи (отношения) между ними;

– правила вывода.

Онтология должна иметь формат, который может обрабатываться компьютерными программами.

Краткое описание классов понятий онтологии ПрО “Оценка и анализ пожаровзрывобезопасности ситуаций на объектах типа ХВП”:

– виды опасности (взрыв, пожар);

– хранилища и их элементы (здания, сооружения, помещения, зоны хранения и т.п.);

– инфраструктура и техническое оборудование для транспортировки, погрузки, специальных работ, противопожарных и охранных действий;

– материалы и предметы хранения;

– персонал хранилища (управление, технический персонал, пожарная охрана, охрана хранилища);

– события, возникающие на ХВП (базисные события (БС), сечения БС, аварийные последовательности, цепочки нежелательных событий, верхние события дерева отказов, исходные и конечные события дерева событий);

– причинные факторы пожаровзрывоопасности. Описание классов причинных факторов приведено в [4];

– ситуации, возможные на ХВП;

– относительные роли факторов (значимости) при возникновении нежелательных событий, приводящих к пожарам и взрывам;

– вероятностные оценки степени опасности ситуаций, возникающих на ХВП;

– мероприятия по предотвращению чрезвычайных происшествий на ХВП (инженерно-технические, организационные, кадровые, противопожарные).

Связи (отношения) между понятиями.

Приведем некоторые отношения, характерные для онтологии ХВП:

а) $< Цеп_{j_i} \rightarrow Ав_i >$,

где $Ав_i$ – авария (ЧП, взрыв, пожар), $i \in I$;

I – множество индексов возможных аварий;

$Цеп_{j_i}$ – аварийная последовательность опасных событий (цепочка), приводящая к $Ав_i$;

$j_i \in J_i$, – множество индексов цепочек, приводящих к $Ав_i$;

б) $< OnC_{\beta} member\ of\ Цеп_{\alpha} >$,

где OnC_{β_α} – опасное событие, являющееся элементом цепочки $Цеп_{\alpha}$;

$\beta_\alpha \in B_\alpha$, B_α – множество индексов OnC из цепочки $Цеп_{\alpha}$.

в) $< C_{l_k} \rightarrow OnC_k >$,

где C_{l_k} – сечение (минимальное множество базисных событий, приводящее к опасному событию) OnC_k;

$l_K \in L_K$, L_K – множество индексов сечений, порождающих OnC_k;

г) $< БС_{\gamma_\delta}\ member\ of\ C_\delta >$,

где $БС_{\gamma_\delta}$ – базисное событие, являющееся элементом сечения C_δ;

$\gamma_\delta \in Y_\delta\ Y_\delta$ – множество индексов БС из сечения C_δ.

д) $< F_{Ф_\gamma}\ member\ of\ БС_\gamma >$,

где $F_{Ф_\gamma}$ – причинный фактор, влияющий на возникновение базисного события $БС_\gamma$;

$ф_\gamma \in Ф_\gamma$, $Ф_\gamma$ – множество индексов факторов, влияющих на возникновение $БС_\gamma$.

3. Технология извлечения знаний из текстовых ИР с использованием онтологии ПрО и семантической разметки

Можно выделить три этапа в данной технологии: подготовительный, семантической разметки и извлечения знаний.

3.1. Подготовительный этап включает следующие шаги

3.1.1. Построение онтологии ПрО

Исходя из целей и сферы применения онтологии, формируются основные понятия ПрО (метаданные). Устанавливаются однозначные термины метаданных и их обозначения [5–8]. Выделяются классы сформированных понятий и определяются связи (отношения) между ними, в том числе отношения “часть/целое”, “элемент/класс”, “подкласс/класс”, “причина/следствие“.

3.1.2. Формирование набора наиболее употребительных лексических выражений, используемых в ПрО

Эксперт выбирает из текстовых материалов ПрО слова и словосочетания, являющиеся устоявшимися терминами и выражениями, принятыми в ПрО. Данную работу целесообразно автоматизировать, используя программное обеспечение, которое позволяет: просматривать электронные тексты из специализированных журналов, диссертаций, словарей, энциклопедий; отмечать характерные слова и словосочетания; фиксировать выбранные элементы в специальном файле “Выборка”.

3.1.3. Формирование тезауруса специальных терминов и понятий ПрО (Тез ПрО)

Здесь главная задача данного шага–установить соответствие между каждым понятием онтологии и словосочетаниями, которые являются синонимами понятия либо близко связаны с ним по смыслу.

В результате все словосочетания, соответствующие одному понятию онтологии, включаются в один класс ПрО.

Формально данное соответствие может быть представлено:

$$\forall C_i, \quad C_i \in C \; \exists W_i = \{w_{1_i}, w_{2_i}, ..., w_{K_i}\},$$

где C – множество понятий онтологии ПрО, W_i – совокупность слов или словосочетаний, выражающих понятие C_i (класс, соответствующий C_i).

Примечание. Создание подобного тезауруса должно поддерживаться программными средствами по группировке словосочетаний и автоматическому кодированию классов.

Подготовительный этап выполняется экспертами в ПрО при участии разработчиков ЭС.

3.2. Семантическая разметка информационных ресурсов

Целью данного этапа является внесение в электронные тексты документов, составленных на естественном языке, формальных признаков метаданных, характеризующих смысловое содержание документов.

Разметке подвергаются ИР, которые являются источниками знаний для пользователей ЭС по ПрО, (электронные версии специализированных журналов по взрывопожарной безопасности, профильные отчеты Министерства по чрезвычайным ситуациям, а также материалы, выявленные пользователем из Интернета при поиске по ключевым словам). Можно выделить следующие шаги данного этапа.

3.2.1. Разбиение ИР на фрагменты

Фрагментирование не является обязательным. Его целесообразно выполнять, когда ИР достаточно велик по объему. Фрагментами ИР могут быть разделы документа, страницы или абзацы.

3.2.2. Первичная семантическая разметка ИР

Для каждого понятия онтологии ПрО C_i в тезаурусе (Тез ПрО) выполняется поиск класса лексических элементов, соответствующих данному понятию: $\{w_{1_i}, w_{2_i}, ..., w_{K_i}\}$. Затем выполняется поиск этих лексических элементов в размечаемом ИР. В случае, если в некотором фрагменте обнаружен хотя бы один элемент из множества $\{w_{1_i}, w_{2_i}, ..., w_{K_i}\}$, то данному фрагменту присваивается «ярлык» (ТЭГ), соответствующий понятию C_i (соответствие между понятиями онтологии ПрО и множеством ТЭГов устанавливается на основании Тез ПрО).

В результате применения подобной процедуры ко всем понятиям онтологии ПрО каждому фрагменту размечаемого текста будет присвоено j ТЭГов, $j = (\overline{O, N})$, где N – количество понятий онтологии.

3.2.3. Вторичная (дополнительная) разметка ИР

На втором этапе выполняется разметка с учетом онтологических отношений между понятиями.

Рассмотрим пример. При построении онтологии ПрО были зарегистрированы понятия: “Перегрузка персонала” (ПП) и “Человеческий фактор как источник опасности” (ЧФ). При этом между ними было установлено и зафиксировано отношение

$$< ПП \ подкласс \ ЧФ >, \qquad (*)$$

т.е. “перегруз персонала (во всех его формах) является подклассом причин опасности, вызванных человеческим фактором”.

Допустим, что некоторый фрагмент размечаемого текста ФрА содержит только лексические единицы, соответствующие понятию “ПП”. Тогда при первичной разметке ему был присвоен ТЭГ “ПП”.

При вторичной разметке было учтено отношение (*) и фрагменту ФрА будет дополнительно присвоен ТЭГ ”ЧФ”.

Первичная и вторичная разметки выполняются автоматически программными средствами. При этом первичная разметка реализуется одним алгоритмом для всех понятий онтологии. Алгоритм вторичной разметки должен содержать столько логических вариантов, сколько типов связей между понятиями онтологии должно быть учтено при разметке.

Результатом семантической разметки является совокупность фрагментов, каждый из которых, наряду с исходными данными, содержит набор ТЭГов, соответствующих понятиям онтологии содержащимся во фрагменте, т.е.

$$размеченный \ \ текст = \{new \ \ frag_l\},$$

где $l = (\overline{1, L})$, L – количество фрагментов текста, $new \ \ frag_l = [(исходный \ \ фрагм.) \& \{T \ э \ \Gamma_{i_l}\}]$,

$i_l = (\overline{1, I_l})$, I_l – количество ТЭГов в l-ом фрагменте.

Размеченные ИР помещаются в библиотеку размеченных текстов, откуда могут заимствоваться для целевого извлечения знаний.

3.4. Извлечение знаний из размеченных текстов

Технология извлечения знаний из текстов аналогична поиску по ключевым словам с тем различием, что в запросах пользователя применяются не ключевые слова, а понятия онтологии ПрО.

Выбор искомых фрагментов выполняется по критерию соответствия запроса пользователя и ТЭГов, описывающих понятийное содержание фрагментов. Конкретные особенности формирования запросов пользователя определяются спецификой информационно-поисковых систем, которые выбраны для извлечения знаний.

Примечание. Пакет текстовых фрагментов, отобранных в процессе поиска, подвергается дальнейшей обработке с целью формализации полученных знаний. Данная процедура реализуется в интерактивном режиме инженером по знаниям. При этом выполняются следующие действия: группировка фрагментов по понятиям онтологии; удаление смысловых повторов; смысловое обобщение содержания отдельных фрагментов; формирование формализованного описания элементов базы знаний (согласно выбранному языку представления знаний); загрузка результатов в базу знаний экспертной системы. Описания этих процедур будут изложены в дальнейших работах.

4. Заключение

Предлагается подход к созданию варианта автоматизированной технологии извлечения знаний из текстовых информационных ресурсов, в котором используется семантическая разметка исходных текстов метаданными узкой предметной области (ПрО).

Семантическая разметка основана на предварительно установленном соответствии лексических выражений текстов и метаданных, описанных в онтологии ПрО.

Семантическая разметка, выполняемая автоматически, значительно повышает возможности формализации понятийного описания фрагментов текстов ПрО и, кроме того, позволяет извлекать знания не по ключевым словам, а с помощью поиска по метаданным.

СПИСОК ЛИТЕРАТУРЫ

1. Рогушина Ю.В. Интеллектуальные поисковые системы в контексте технологий Semantic Web / Ю.В. Рогушина, А.Я. Гладун // Интеллектуальный анализ информации. – 2008. – № 2. – С. 388 – 398.
2. Semantic Wiki [Електронний ресурс]. – Режим доступу: http://en.wikipedia.org/wiki/SemanticWiki.
3. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб: Питер, 2000. – 384 с.
4. Серебровский А.Н. Подход к созданию базы знаний экспертной системы оценки прогноза и анализа ситуаций на объектах повышенной опасности / А.Н. Серебровский // Математичні машини і системи. – 2009. – № 4. – С. 58 – 66.
5. Система предупреждения и действий в чрезвычайных ситуациях. Понятийно-технологический словарь. – Минск: Полыня, 1992. – 311 с.
6. ГОСТ 12.1044. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения. – Введ. 01.01.91; Переизд. 01.01.96 [Электронный ресурс]. – Режим доступа: http://www.fireman.ru/bd/gost/12-1-044-89/12-1-044.html.
7. ДСТУ 2272-93. Пожежна безпека. Терміни і визначення. ССБТ. – Введ. 15.02.09. – 8 с.
8. ГОСТ 12.1044-91. Пожарная безопасность. Общие требования. – Введ. 01.07.92 [Электронный ресурс]. – Режим доступа: http://www.fireman.ru/bd/gost/12-1-044-89/12-1-044.html.

Стаття надійшла до редакції 24.12.2009