УДК 681.3: 612.84

А.А. Стенин, Ю.А. Тимошин, В.Г. Галаган, В.П. Ярченко

О ФУНКЦИОНАЛЬНОСТИ БАЗЫ ЗНАНИЙ ПО АНАЛИЗУ И ОБРАБОТКЕ ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ

Аннотация: Рассматриваются вопросы формирования информационной среды базы знаний для лингвистического анализа и обработки с такими формами представления, как данные, информация, тематические знания, интеллектуальные знания и опыт применения с использованием метаданных. Дается их описание с использованием функциональных слоев для каждой формы и показываются их взаимосвязи, которые приводят к многоуровневой модели БЗ.

Ключевые слова: база знаний, метаданные, метамодель, семантико-синтаксический анализ, рекурсивная модель обработки

Введение

Знание является сложным, комплексным понятием и может описываться и классифицироваться множеством различных способов. До сих пор нет однозначно принятого определения – что такое знание.

Под знаниями, обычно, понимают набор фактов и правил, формализующих опыт специалистов в конкретной предметной области.

Что касается информации, то и здесь нет однозначности в ее определении и разные авторы [1–3] приводят свои определения и аргументы, поэтому этот вопрос также требует соответствующей определенности.

Процесс управления знаниями включает в себя поиск, сбор, кодификацию, хранение, распространение, а также обновление знаний. Управление знаниями (Knowledge Management) является сложной задачей для многих компаний. И именно технологии играют ключевую роль при внедрении конкретного подхода. Сегодня существует множество решений в области управления знаниями с различной функциональностью. Это SAP Knowledge Management (SAP KM), MS SharePoint , Domino Document Manager, Hummingbird KM, IBM Information Management Software, Documentum, Media Wiki, другие. Готовое решение, безусловно, может помочь при создании компонентов базы знаний (БЗ), но не решит всех проблем с формированием БЗ.

Часто базу знаний рассматривают как один из типовых компонентов или системы поддержки принятия решений в какой-то узкой проблемной области или другой системы искусственного интеллекта, включающей “решатель”, с системой вывода новых знаний и решений на основе декларативных и др. знаний, соответствующих правил и процедур вывода, а также с БД фактов, процедур и т.д. В статье рассматривается структура БЗ с функциональными формами представления для решения задач лингвистического анализа и обработки, которая отличается от традиционных решений.

 ISSN 1560-8956

Постановка задачи

Формирование базы знаний – многоэтапный сложный процесс. В первую очередь следует выделить основные принципы, которые должны быть заложены в разработку форм представления, описания и формирования наполнения БЗ информационными объектами, знаниями о технологиях и схемах их хранения, анализа и обработки, а также по системе управления самой базой знаний. Для задач лингвистического анализа и обработки основными будем считать задачи морфологического разбора слов и речи, задачи синтаксического анализа предложения и семантического анализа текста, а также тематические знания по их моделям, опыту применения и использования.

Кроме данных, информации, тематических знаний, интеллектуальных знаний по их использованию в нашей БЗ должны храниться и обрабатываться также объекты/субъекты/ процессы/явления/события/ситуации, которые выявляются при обработке исходных текстов в виде лингвистических объектов и которые имеют различного вида связи и отношения между собой, в том числе синтаксические, семантические и ассоциативные. Учитывая большое разнообразие типов объектов и связей, которые выявляются в результате лингвистического анализа в полипредикатных и монопредикатных структурах текстов, необходимо осуществлять обработку огромного числа возможных вариантов структурных связей и получаемых структур этих текстов. В [5] предложена технология и модель Базовой семантико-синтаксической структуры (БССС), которая позволяет частично сократить число переборов вариантов для сложных предложений полипредикатной структуры за счет рекурсивной организации процесса анализа лингвистических объектов.

Использование метамоделей для описания, хранения и обработки данных, информации и знаний позволяет создать несколько слоев абстракции и значительно сократить объем перерабатываемой информации и данных в задачах лингвистической обработки [4] текстов с монопредикатной структурой.

Существует также проблема измерения или оценивания в БЗ величины информации, знаний и интеллекта [6], для решения чего надо разработать соответствующие метрики, которые представляли бы ясные и четкие величины для практического использования как специалистами по обработке знаний, так и специалистами по лингвистическому анализу.

В данной статье мы будем придерживаться точки зрения, что информация, которая есть в БЗ, состоит из данных и метаданных для их описания, идентификации, хранения и обработки, знания в БЗ - это информация и описание ее назначения (смысл, толкование), интеллект – это знания и умение их использовать для решения задач по лингвистическому анализу. Есть еще одно измерение, которое может быть использовано в среде БЗ – это описание опыта (или навыков) реализации ин-

теллектуальных знаний. Т.о. можно получить пятимерное представление и описание в БЗ тех лингвистических объектов, которые могут в ней храниться и обрабатываться.

Кроме указанных объектов и компонентов, БЗ включает различные средства хранения (базы данных, репозитории, каталоги, тематические справочники и классификаторы), средства обработки и анализа/выявления новых фактов и знаний, др. системы, которые образуют систему поддержки БЗ.

Указанные выше объекты, задачи и компоненты определяют, в основном, функциональность описываемой БЗ по лингвистическому анализу и обработке.

Параметры и характеристики функциональных слоев

Рассмотрим, какие параметры и характеристики составляют каждый слой БЗ.

Для слоя данных можно определить следующие компоненты: исходные текстовые файлы, символы, даты и периоды, лингвистический контент (слова, словосочетания, словесные обороты, предложения, абзацы, страницы, текст и речь), мультимедиа данные (аудиоряд, видео- ряд, изображения и фото, видео- фильм), документы и файлы различного вида (включая рисунки и схемы, алгоритмы и процедуры), лингвистические объекты, которые распознаются в текстах (объекты/субъекты/процессы/явления/состояния/события), а также запросы к БЗ и фактические результаты анализа и обработки.

К слою информации можно отнести следующие характеристики и метаданные: представления о лингвистических и физических объектах, параметры систем обработки, словари и справочники, классификаторы, источники данных (параметры и характеристики, интерфейсы и драйвера, шаблоны доступа и запросов), формы (регистрации, контроля, представления и отображения), шаблоны информации и структуры данных, метаданные (описания, хранения, представления и управления), идентификаторы, ключи и индексы, ссылки, языки и коды, контрольная сумма, описание сервисов, процедур анализа и обработки (назначение, исходные данные и параметры, среда обработки), показатели видов связи и отношений, характеристики процессов анализа и обработки (например, затраты времени, памяти, идентификатор *.dll-библиотеки, IP-адрес порта, т.п.), свойства лингвистических объектов (морфологические, синтаксические, семантические), каталог и тезаурус БЗ (в системах общего доступа они могут быть отнесены к слою тематических знаний по мнению авторов).

К слою тематических знаний предлагается отнести такие характеристики, которые позволяют определить назначение, применение, содержание, значение/цель или толкование предмета знания, смысла этапа или представления (объекта/субъекта/процесса/явления/состояния /события):

 ISSN 1560-8956

a) Понятия, на основе которых формируются представления и наборы из них для создания информационного обеспечения по лингвистическим объектам;

b) Наборы представлений для объектов/субъектов/процессов/явлений/событий/ситуаций;

c) Связи и отношения в среде лингвистических объектов, совместно с наборами представлений, в моделях и метамоделях;

d) Метамодели, которые описывают метаданные предыдущего слоя;

e) Алгоритмы анализа и обработки (в том числе и лингвистического назначения), их описания, ограничения и начальные условия;

f) Модели разного назначения, которые используются в БЗ и в задачах анализа и обработки;

g) Описания технологий и процессов анализа и обработки;

h) Параметры управления, их диапазоны и ограничения;

i) Критерии управления и метрики, на которых построен реальный критерий (останова, итерации или передачи управления, точности обработки и т.п.);

j) Идентификаторы объектов/субъектов/процессов/явлений/событий/ситуаций;

k) Этапы обработки или анализа, последовательность этих этапов и описание особенностей, порядок их контроля;

l) Шаблоны обработки и анализа (например, шаблоны запросов к БЗ, шаблоны описаний лингвистических объектов, т.п.);

m) Описания форм представления информации;

n) Описания задач анализа или обработки;

o) Модель или описания желаемого результата или цели анализа /обработки (в разрезе параметров или характеристик, которые можно измерить или вычислить);

p) Программное и системное обеспечение (например, используемые модули приложений и СУБД);

q) Описание средств шифрования/кодирования/доступа;

r) Каталог метамоделей описания лингвистических объектов;

s) Реестры запросов и источников данных (в системах общего доступа они могут быть отнесены к слою информации, по мнению авторов).

К слою интеллектуальных знаний можно отнести знания и умения по использованию тематических знаний и по их обработке средствами поддержки БЗ следующего вида:

a) Особенности выполнения лингвистического анализа и обработки (морфологический, синтаксический, семантический);

b) Критерии классификации или идентификации в зачах лингвистического анализа или обработки;

c) Выполнение процессов классификации или идентификации в зачах анализа или обработки;

d) Создание наборов компьютерных представлений о лингвистических объектах;

e) Правила связывания компьютерных представлений и соответствующих лингвистических объектов;

f) Процедуры выбора любого критерия или метрики в задачах;

g) Процесс построения любого критерия;

h) Идентификация связей и отношений в лингвистических моделях текста или в метамоделях;

i) Выбор формы представления результатов или формы документа о результатах;

j) Правила и порядок учета условий, ограничений и предписаний (например, формата или прикладного протокола);

k) Выбор схемы алгоритма или процесса;

l) Выбор правил обработки или анализа;

m) Выбор любого шаблона для обработки или анализа;

n) Процессы построения или уточнения алгоритма или схемы анализа и обработки;

o) Конкретизация последовательности этапов анализа и обработки;

p) Все действия по интеграции процессов и алгоритмов;

q) Процесс выработки решения;

r) Процесс принятия решения;

s) Процесс оценки результатов лингвистического анализа и обработки;

t) Процесс построения модели “желаемого результата”;

u) Процессы анализа фактов и выявления новых знаний;

v) Процессы формирования новых знаний и их применения;

w) Оценка признаков “ценности” информации или знания;

x) Согласование цели и конечного результата анализа и обработки;

y) Выбор прикладного ПО для решения задачи или задания;

z) Каталог метамоделей хранения и управления обработкой лингвистических объектов.

К слою описания опыта (или навыков) реализации интеллектуальных знаний

 ISSN 1560-8956

a) Использование опыта и примеров положительных решений конкретных задач лингвистического анализа и обработки;

b) Использование примеров “неудачных” решений задач;

c) Варианты решения прошлых задач, их условия и ограничения;

d) Старые модели и шаблоны;

e) Метрики и критерии, использованные в старых задачах;

f) Сравнительный анализ

g) Результаты и выводы старых проектов и задач в соответствующем реестре БЗ;

h) Представления и их объекты в задачах;

i) Наборы представлений (образцы), связанных с объектами в старых задачах;

j) Схемы обработки и анализа;

k) Способы индексации в старых задачах;

l) Справочные описания объектов, процессов, ситуаций и решений из старых задач;

m) Алгоритмы шифрования и процедуры доступа;

n) Обобщения, резюме и выводы экспертов;

o) Архивы представлений, документов и метамоделей

p) Репозитарии хранилища и БЗ;

q) Список ссылок с URL-адресами;

r) История развития БЗ (средства журнализации).

Описанные выше параметры и характеристики сведены и представлены в Табл. 1.

Для этих функциональных слоев необходимо далее разработать соответствующие метрики прямого измерения или вычисления указанных компонентов, позволяющих спроектировать различные критерии для дальнейшей оптимизации обработки и анализа лингвистических объектов, что будет рассмотрено в следующих статьях.

Таблица 1

Функциональные слои базы знаний

Одномерные	Двухмерные	Трёхмерные	Четырёхмерные	Пятимерные
Данные	Информация (данные + метаданные)	Знания (информация + смысл)	Интеллект (использование знания + умения)	Опыт (навыки решения задач, реализация интеллекта)
+исходн. данные +факт. результаты +символы +файлы +контент (слова/ предложения /абзацы/ страницы/ текст) +видеоряд +изображения +аудио-ряд + multi-media data +объекты анализа или обработка	+представления (объектов, событий, ситуаций) +параметры обработки +интерфейсы и средства распаковки +словари +справочники +источники данных +параметры источников д. +драйвера +*dll-библиотеки +форматы распаковки +характеристики/ свойства +шаблоны представления +время анализа/ обработки +метаданные описания структуры данных +идентификаторы +форматы представления и отображения +языки/ коды +входы и выходы +ключи/ индексы/ ссылки +сервисы +показатель связи / отношения +контрольная сумма +даты/ периоды +средства описания/ оценивания /распознавания + значение ценности информ. +оценки желаемого результата	+понятия +наборы представлений +метамодели +алгоритмы +ограничения +условия +модели +технологии +процессы +параметры управления +критерии (метрики) +категории, виды, классы +классификаторы +правила обработки +средства шифрования (кодирование/ декодирование) +этапы/ схемы обработки +шаблоны обработки и анализа +порядок контроля анализа/ обработки результатов +программные средства (ПО И СУБД) +виды и типы связей и отношений +описание форм представления и визуализации информации +типовые задачи анализа или обработки +желаемый результат +показатели и оценки ценности информации	+выполнение лингвистического анализа и обработки (морфологический, синтаксический, семантический) +выполнение классификации +создание набора представлений +выбор критерия +учёт условий, ограничений и предписаний +выбор алгоритма +выбор шаблона +выбор и конкретизация последовательности этапов обработки или схемы обработки +критерии выполнения процессов анализа и обработки +выработка решений +принятие решений +все действия по интеграции +выбор форм представления +оценка ценности информации +выполнение оценивания близости реального и желаемого результатов +разработка новых или доработка старых задач анализа или обработки +выбор стратегий поиска лингвистических объектов , их образов и представлений +выявление новых фактов и знаний, их согласование, описание и формирование в БЗ	+удачные/ неудачные решения задач +варианты решения, их условия и ограничения задач +новые решения и модели +старые решения и модели +новые схемы, алгоритмы и шаблоны (обработки/ анализа) +новые и старые метрики и критерии +новые парадигмы +драйвера/ интерфейсы +любой сравнительный анализ +результаты и выводы старых проектов и задач +представление и их объекты +способы идентификации +набор представлений (образцы) +схемы обработки и анализа +справочные описания (объектов/ процессов/ ситуаций/ решений/ задач) +Help(ы) приложений +алгоритмы шифрования +резюме и обобщения + репозитарий хранилища и список ссылок с URLами +история развития БЗ

ISSN 1560-8956

О режимах функционирования БЗ

Описанные функцональные слои обеспечивают функционирование такой БЗ в трех основных режимах работы:

- “Режим обучения БЗ” - когда производится обновление справочников, классификато-ров, репозитариев, тезауруса или критериев анализа и обработки БЗ, включая “залповый” ввод текстов, что может выполняться практически без участия челевека, или с участием эксперта в интерактивном диалоговом режиме;
- “Режим анализа и обработки” - когда выполняются типовые процессы лингвистичес-кого анализа и обработки для исходных текстов с целью их дальнейшей структуриза-ции в виде метамоделей, определения представлений для лингвистических объектов, формирования связанных наборов представлений для уточнения выполняемого анализа, т.п., а также при решении различных неформализованных задач обработки;
- “Режим обнаружения новых фактов и знаний” - когда в результате лингвистического анализа или обработки выявляются новые факты или результаты, уточняются условия и ограничения применения алгоритмов и процедур, схем и процессов, на основании чего экспертом могут быть сформированы новые знания для вашей БЗ.

Выводы

В пределах принятых авторами ограничений относительно функциональной трактовки информации и знаний, показана возможность формирования различных моделей представления информации и знаний для лингвистических приложений. Учитывая необходимость анализа функциональности БЗ и поиска новых знаний в 4-х или 5-и мерном пространстве, возникает дополнительно задача разработки специальных программных средств для такого анализа с визуализацией получаемых на каждой итерации результатов.

В следующих выпусках сборника авторы надеются опубликовать материалы по разработанным измеримым метрикам для указанных слоев и по архитектуре БЗ и обслуживающих ее систем лингвистического анализа и обработки, разработанной с использованием описанной выше функциональности.

Авторы осознают сложность затронутых в статье вопросов формирования базы знаний, а также широкий спектр проблем, связанных с этим, и поэтому не претендуют на прямое обобщение предложенного подхода для других тематических задач анализа, не связанных с описанной предметной областью.

Литература

1. Кузнецов С.В. Технологии управления, основанного на знаниях// Проблемы теории и практики управления (Москва). – 24.12.2004.- 006.- с. 85-89

2. В.А. Молчанова, Т.А. Тумина. О формировании интегрированной базы знаний компании.–//Транспортное дело России, 2008.

3. Питер Джексон Введение в экспертные системы = Introduction to Expert Systems — 3-е изд. — М.: “Вильямс, 2001. — С. 624. —ISBN 0-201-87686-8

4. Шемсединов Т.Г. Слой ИС с динамической интерпретацией метаданных. http://blog.meta-systems.com.ua/2011/01/blog-post_28.html

5. Кисленко Ю.І. Інформаційний підхід до аналізу структурного рівня мовної організації. В сб. трудов междунар. конф. “Языковые технологии в современном мире”, Ялта, 2010, стр. 90-101

6. Джозеф Джарратано, Гари Райли “Экспертные системы: принципы разработки и программирование” : Пер. с англ.— М.: Издательский дом “Вильямс”, 2006.—1152 с.

Отримано 06.03.2012 р.