УДК 681.3.07

О.В. Золотарёв[1]
Е.Б. Козеренко[2]
М.М. Шарнин[3]

ПРИНЦИПЫ ПОСТРОЕНИЯ МОДЕЛЕЙ БИЗНЕС-ПРОЦЕССОВ ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ ОБРАБОТКИ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

В статье обсуждаются методы конструирования моделей предметной области, основанной на извлечении объектов и процессов из текстов естественного языка. Рассматриваются подходы к созданию базы знаний на основе механизма расширенных семантических сетей. Кроме этого, в статье описываются принципы построения и расширения терминологических тезаурусов.

Ключевые слова: *текст естественного знания, бизнес-процесс, семантические сети, тезаурусы, ключевые слова, предметная область.*

O.V. Zolotarev
E.B. Kozerenko
M.M. Sharnin

PRINCIPLES OF BUSINESS PROCESSES MODELS CONSTRUCTION OF SUBJECT DOMAIN ON THE BASIS OF PROCESSING OF NATURAL LANGUAGE TEXTS

In the article is discussed methods of constructing the models of the subject domain based on extracting objects and processes from the natural language texts. It is discussed approaches to creating the knowledge base with the mechanism of extended semantic networks. Besides the article describes principles of creating and adding of term thesauruses.

Keywords: *natural language text, knowledge, business process, semantic networks, thesaurus, keywords, subject domain.*

Введение

В настоящее время весьма актуальной является задача повышения эффективности деятельности предприятия вследствие растущей конкуренции среди различных предприятий не только на мировом рынке, но и внутри России. Особенно это касается предприятий с устаревшими методами работы, с давно устоявшимися бизнес-процессами. В первую очередь в улучшении работы предприятия заинтересованы его владельцы, потому что они получают основную часть прибыли по результатам его функционирования, затем идет руководство предприятия и, в конечном итоге, рядовые работники. Первоначально для выяснения причин недостаточно эффективной работы организации необходимо провести анализ ее деятельности с целью поиска проблем, несоответствий в работе предприятия, излишних затрат и т.д. В результате анализа формируются формализованные описания бизнес-процессов, на основе которых в дальнейшем будет построена модель текущей деятельности предприятия. Эта модель подвергается критическому анализу, в результате которого строится модель будущих процессов.

В рамках описанных процессов одним из самых затратных является процесс построения модели текущих бизнес-процессов деятельности предприятия. Поэтому весьма актуальным является вопрос выделения из текстов естественного языка бизнес-процессов для упрощения процесса проектирования, к которому и относится этап построения модели текущих бизнес-процессов.

1. Анализ мировой практики извлечения бизнес-процессов из текстов естественного языка

Для выделения бизнес-процессов из текста необходимо произвести морфологический разбор анализируемых текстов с выделением харак-

[1] Кандидат технических наук, доцент, доцент НОУ ВПО «Российский новый университет».

[2] Кандидат филологических наук ИПИ РАН.

[3] Кандидат филологических наук ИПИ РАН.

терных объектов предметной области, связанных с ними свойств, связей между объектами и действий – процессов, в которых задействованы выделенные объекты.

Существует достаточно большое количество подходов в мировой практике по выделению бизнес-процессов из текста.

Во многих публикациях, относящихся к данной проблеме, анализируются достаточно большие объемы текстовой информации, используемой для анализа с целью выделения объектов и процессов, например система Rocket AeroText.

Предлагаются различные подходы для построения бизнес-процессов на основе анализа информации их устаревших систем (Nascimento (2012)). Эта проблема стоит особенно остро, когда необходимо не просто переносить информацию их из старых информационных систем в новую систему, но и когда приходится ее преобразовывать, структурировать, извлекать из нее объекты, свойства, связи и процессы, строить на основе извлеченной информации бизнес-правила.

В подходе, предложенном в 2008 году Putrycz and Kark предлагается строить бизнес-правила на основе анализа естественно-языковых текстов в формате ЕСЛИ.., ТО... При выполнении определенного условия выполняется определенное действие.

Важно уметь не только выделять объекты и процессы из текста, но и структурировать информацию на основе анализа больших объемов текстовой информации, что является весьма затруднительным, потому что выделить и идентифицировать объекты и связанные с ними действия на основе анализа одного предложения не так сложно. Значительно соотнести выделенные во всем тексте объекты и процессы с уже найденными, выстроить их взаимозависимости, понять, как они взаимодействуют друг с другом, соотнести их к конкретной ситуации, которая также должна быть идентифицирована. Известны работы Chapparro et al. (2012) в рамках исследовательского проекта. Предлагается использовать шаблоны структурированных конструкций правил. При этом в качестве источников информации рассматриваются различные базы данных, текстовая информация, документы, документация к различным системам, электронная почта и т.д.

Подобные работы ведутся по всему миру. Конечно, наиболее проработанными считаются подходы, связанные с анализом англоязычных тестов. Себастьян Падо (Германия) и Мирелла Лапата (Великобритания) строят семантические пространства на основе статистического подхода с формированием векторных моделей. Выделяются целевые слова, строится семантическая матрица. Каждое слово – это точка в многомерном пространстве. Определяется семантическое сходство, понятие близости терминов.

2. Моделирование деятельности предприятия

В настоящее время в целях повышения конкурентоспособности на многих предприятиях внедряются информационные системы (ИС) и корпоративные ИС (КИС), что позволяет в существенной степени упорядочить бизнес-процессы (БП) предприятия, повысить эффективность его деятельности, сократить себестоимость продукции, повысить ее качество и, соответственно, занять или расширить свою нишу на рынке. На многих предприятиях выполняются проекты по реинжинирингу бизнес-процессов. При внедрении КИС этап реинжиниринга БП является обязательным, потому что нельзя автоматизировать «хаос», и, кроме этого, КИС всегда налагает дополнительные требования на описание и структуру БП предприятия. Одним из наиболее трудоемких этапов внедрения КИС является процесс проведения обследования и формирования требований к будущей системе, а также процесс проектирования, включающий построение статических и динамических моделей деятельности предприятия. Именно обсуждению оптимизации описанных выше этапов и посвящена данная статья.

В процессе обследования предприятия изучается большое количество документов, проводятся опросы, анкетирование, анализируется документооборот предприятия и т.д. Во многом данный процесс можно ускорить и упростить на основе автоматизированной обработки текстовых документов и построения в результате моделей деятельности предприятия, которые затем лягут в основу работ по проектам совершенствования деятельности организации.

В данном контексте особое значение приобретают средства по автоматическому извлечению информации из текстов естественного языка, которые позволят не только определить объекты, их характеристики и отношения, но и связывающие эти объекты действия – процессы. В результате можно будет построить статические и динамические модели деятельности предприятия, выполнив тем самым обязательный этап процесса внедрения КИС или реинжиниринга бизнес-процессов (РБП).

3. Инструментальные средства построения моделей предметных областей

Сегодня на рынке представлено достаточно большое количество средств проектирования, ориентированных на построение компьютерных моделей предметной области.

Одно из средств поддержки формата – семейства IDEF – ERwin – Bpwin. В рамках этой системы реализована возможность построения диаграмм в форматах IDEF0, IDEF3, DFD. На основе формата IDEF3 может быть построена диаграмма типа Swim Lane. Инструментальная среда Bpwin предназначена для описания бизнес-процессов или для построения динамической стуктуры ПО, ERwin – для построения статической структуры ПО.

MS Visio поддерживает разработку диаграмм и блок-схем различного назначения: аудит, дерево ошибок, организационные диаграммы, причинно-следственные связи, диаграммы маркетинга, карты вычислительных сетей, каталогов LDAP и Active Directory, карты сайтов, связи между объектами в программном обеспечении, структуры и интерфейсы программ, потоки данных, планы помещений, этажей, инженерно-технических коммуникаций, схемы рабочего процесса, чертежи и схемы электронных устройств и т.д. Также поддерживаются форматы IDEF0, IDEF3, DFD, язык UML.

Business Studio – система бизнес-моделирования разработана на базе MS Visio 2003 и предназначена для описания бизнес-процессов предприятия, автоматизации управления внутренней нормативной документации, описания организационной структуры предприятия.

Система QPR позволяет не только описывать бизнес-процессы предприятия, но и анализировать его деятельность с помощью системы сбалансированных показателей, управленческих и операционных показателей, находить узкие места бизнес-процессов. Также поддерживаются форматы IDEF0, IDEF3, DFD.

ARIS – один из наиболее мощных инструментов для моделирования бизнес-процессов предметной области. Это не только программный продукт, но и методология. В ARIS поддерживается и моделирование данных, и моделирование бизнес-процессов. Также ARIS поддерживает язык UML. Отличительная особенность ARIS – возможность создания сценариев составления отчетов, аналитических документов, моделей на специальном языке. Продукт используется в проектах SAP и ORACLE.

Кроме этого существует большое количество сред для поддержки языка моделирования UML – Unified Modeling Language, объектно ориентированный язык моделирования, например IBM Rational Rose, StarUML и др. В UML могут быть построены различные типы статических и динамических диаграмм на основе объектно ориентированного подхода.

4. Статистические методы обработки текстов естественного языка

Для представления знаний, извлекаемых из текста, существует широкий спектр различных средств. В данной статье рассматривается подход, основанный на семантических сетях. Это довольно удобный механизм для описания знаний, извлекаемых из текстов естественного языка. На первом этапе проводится морфологический и синтаксический анализ текстов, нормализуются слова, формируются словосочетания. Это объекты, которые могут вступать в отношения с другими объектами, формировать группы. Одна из важнейших проблем при анализе текстов естественного языка – соотнесение текста с рубрикой или определение предметной области, к которой относится текст. Для этого необходимо отделить незначащие слова, т.е. общеупотребительные. Обычно эта процедура выполняется на основе статистического анализа. Незначащие слова – это те, частота встречаемости которых достаточно высока. Задача соотнесения текста с конкретной предметной областью решается на основе классификаторов-тезаурусов, которые каждый раз при обработке нового текста автоматически достраиваются. Тезаурусы проверяются экспертами на предмет корректности заполнения. Для первоначального построения тезаурусов использовались классификаторы. На следующем этапе анализируется окружение объектов с целью выделения характерных для объектов свойств или атрибутов. Каждый раз при этом происходит модификация тезаурусов. После завершения работы с отдельными объектами, их связями и группами выделяются процессы, в которых объекты могут быть задействованы. Цель данного этапа – выделение бизнес-процессов из текста.

Для выполнения анализа тексты могут выбираться из глобальной сети Интернет. При проведении анализа текстов может решаться сопутствующая задача идентификации интересов интернет-пользователей на основе анализа обращений пользователей к различным ресурсам.

Как уже говорилось, соотнесение текста с конкретной предметной областью происходит на основе сопоставления терминов с ключевыми словами. Именно ключевые слова являются

основными элементами для формирования тезаурусов. Ключевые слова могут относиться к различным категориям и могут иметь вес, который определяет значимость термина. Категории представляются следующим образом в виде кортежа:

<Первичное ключевое слово, Категория, Вес>.

При этом оценивается вес каждого термина экспертами, которые и проверяют правильность соотнесения термина к конкретной ПО.

В этом случае может быть посчитана вероятность правильного решения эксперта.

$Pi = p$ (Категория | Слово),

где Pi – вероятность корректного решения i-го эксперта.

Вес каждого эксперта вычисляется по формуле:

$Wi = \log(Pi/1\text{-}Pi)$.

Попутно решается задача анализа предпочтений пользователей Интернета по их запросам. В результате определения рубрик анализируемых текстов могут формироваться портреты интернет-пользователей.

Достоверность приведенных расчетов подтверждается практикой, причем результаты расчетов достаточно устойчивы и не зависят от размера обрабатываемых текстов.

5. Построение модели предметной области с использованием аппарата семантических сетей на основе анализа текстов естественного языка

В рамках описания моделей понятия «процесс» и «подпроцесс» рассматриваются как взаимозаменяемые. Эти понятия различаются только при рассмотрении процессов разных уровней (процесс – более верхний уровень, или родитель, подпроцесс – подчиненный уровень – потомок).

На этапе морфологического анализа текстов строится формализованная структура предложений, выделяются объекты, их свойства, связывающие эти объекты действия. Далее выделяются словосочетания, проводится статистический анализ встречаемости терминов, словосочетаний, действий в тексте. На следующем шаге производится идентификация выделенных элементов в масштабе документа, строится ассоциативный портрет предметной области на основе элементов, извлеченных из документа. В результате анализа представительного корпуса текстов определяется степень семантической близости различных документов. В результате семантически близкие документы соотносятся с конкретной предметной областью. По данной предметной области формируется библиотека документов. На основе статистического анализа портретов документов определяются семантически близкие процессы и объекты. Объекты и процессы представляются в виде семантической сети.

Рассмотрим на примере процесс формирования фрагмента семантической сети на основе анализа текста естественного языка с выделением объектов и процессов.

Далее приводится фрагмент текста, на основе которого строится несколько фрагментов семантической сети. При этом выбрана часть текста, которая является минимальной и достаточной для формирования представленных ниже фрагментов семантической сети, описывающих процесс изготовления детали из заготовки:

«Товар закупается клиентом на основе заказа в соответствии с регламентом. Договор составляется по заказу. После оформления договора производится оплата товара. Товар забирает покупатель».

(1) Закупка (Процесс, _)
(2) Заказ (Объект, Вход, Закупка)
(3) Товар (Объект, Выход, Закупка)
(4) Регламент (Объект, Управление, Закупка)
(5) Клиент(Объект, Участник, Закупка)

В построенном фрагменте (1) определен процесс «…закупается…», на основе которого строится в виде отглагольного существительного действие «Закупка».

Определенный на основе анализа текста процесс «Закупка» располагается во фрагменте семантической сети на 0-м аргументном месте – перед открывающейся скобкой. Сразу же после скобки располагается 1-е аргументное место, которое определяет тип данного элемента. Элемент может быть двух типов: объект или процесс. В данном случае это процесс, причем этот фрагмент является родительским процессом верхнего уровня. На втором аргументном месте фиксируется процесс верхнего уровня, т.е., если у данного родителя в свою очередь есть свой родитель, то 2-е аргументное место будет означено названием процесса более верхнего уровня. Но в нашем случае у представленного фрагмента нет родительской вершины.

На втором аргументном месте у фрагмента, описывающего объект (2), фиксируется вид объекта. В случае если 2-е аргументное место означено (не пустое), то оно указывает на процесс, которому принадлежит данный объект.

Модели процессов могут декомпозироваться (разбиваться на подпроцессы). Процесс «Изготовление» разбивается на два подпроцесса

«Подготовка» и «Вытачивание». В результате будут построены следующие фрагменты семантической сети:
(6) Оформление_договора (Процесс, Закупка)
(7) Оплата (Процесс, Закупка)
(8) Заказ (Объект, Вход, Оформление договора)
(9) Договор (Объект, Выход, Оформление договора)
(10) Договор (Объект, Вход, Оплата)
(11) Товар (Объект, Выход, Оплата)

В результате анализа в документе определено на основе совпадения входов и выходов, а также с учетом временного ряда, какой процесс выполняется вслед за каким. Процессы (6)–(7). Чтобы упростить задачу, рассматриваются только объекты, относящиеся к входам и выходам. Фрагменты, описанные далее (8)–(11), представляют аналогичную часть семантической сети, построенной по аналогии с фрагментами (6)–(7).

Для упрощения отслеживания связей фрагментов вводится дополнительный тип фрагмента, который в явном виде будет указывать на то, каким образом и в какой последовательности фрагменты связаны между собой. В данном случае мы имеем только один подобный фрагмент:
(12) Связь (Оформление договора, Оплата).

Этот фрагмент как раз указывает на последовательность выполнения фрагментов. Фрагменты типа «Связь» – необходимая избыточность, которая позволяет явным образом представлять связи между фрагментами (в неявном виде они могут присутствовать в виде совпадений входов и выходов процессов).

Еще одна возможность определения последовательности выполнения процессов – анализ временного окружения процессов, объектов, их связей. По умолчанию можно говорить о том, что в процессе написания документа временная ось проходит в прямом направлении от начала документа к его завершению. Однако при этом могут использоваться и другие различные способы определения последовательности процессов. Но все-таки главное, на основе чего выстраивается временная цепочка, – последовательное появление процессов в тексте, которое и определяет естественный ход события при выделении объектов и процессов. При этом время выполнения процессов может быть либо указано конкретно, лило вытекать из явного описания последовательности действий. Кроме этого можно обращаться к тезаурусам понятий, в которых также в явном виде могут быть описаны отдельные элементы регламентов выполнения процессов с учетом генерирования стандартных или нестандартных событий.

Построение семантической сети, которая и являет собой базу знаний, основывается на совпадении входов и выходов процессов. При этом надо учитывать связь между уровнями декомпозиции процессов. При разбиении процесса на несколько подпроцессов наблюдается устойчивая связь между входом родительского процесса и входом первого дочернего подпроцесса, а также между выходом родительского процесса и выходом последнего дочернего подпроцесса. Как уже упоминалось, связь между дочерними подпроцессами определяется на основе совпадения входов и выходов дочерних подпроцессов. В результате проведения анализа текстов естественного языка и автоматического построения семантической сети формируются тезаурусы объектов и процессов, а также базы знаний по конкретной предметной области.

Принятие решения об отнесении процесса к определенному уровню может быть принято на основе статистических данных о встречаемости процесса в документе. Чем чаще встречается описание процесса, тем больше вероятность того, что данный процесс является процессом более высокого уровня.

Формирование тезаурусов может выполняться на автоматической и автоматизированной основе. В первом случае возникающие ошибки могут быть исправлены на основе анализа структуры существующих тезаурусов и правил формирования тезаурусов. Во втором случае может быть привлечен эксперт для принятия решения о правильности построения тезауруса. В этом случае могут вноситься правки на основе мнения эксперта. После этого необходимо подтвердить правильность формирования тезауруса, в результате чего будут изменены отдельные правила базы знаний, отвечающие за построение тезаурусов.

В процессе обработки представительного корпуса текстов естественного языка, уточнения и расширения базы знаний количество ручного труда будет сокращаться в существенной степени. Это позволит практически в автоматическом режиме формировать модели бизнес-процессов предметной области на основе накопленных знаний.

Заключение

В данной статье рассмотрен подход к построению моделей предметных областей на основе анализа текстов естественного языка с извлечением объектов, их свойств процессов, в

которых они принимают участие из различного рода документов. В результате обработки текстов выделяются ключевые слова, рассчитывается статистика их встречаемости в документе, строятся тезаурусы понятий. Решается одна из важнейших задач анализа текстов естественного языка – соотнесение документа с конкретной предметной областью, т.е. задача классификации. В процессе использования данного инструмента будут накапливаться знания по различным предметным областям, что позволит в дальнейшем вести обработку текстов практически в автоматическом режиме.

В данной работе представлена концептуальная модель системы по извлечению процессов из текстов естественного языка, основанная на формировании тезаурусов и построении базы знаний объектов и процессов. В результате обработки текстов естественного языка, описывающих некоторую предметную область, строятся наборы фрагментов семантической сети, которые образуют базу знаний данной предметной области. Последующая обработка поступающих документов позволит не только распознать уже выделенные объекты, но и определить их свойства, связи с другими объектами данной предметной области. Наличие большого объема выделенных и выверенных знаний позволяет свести к минимуму количество ошибок при построении функциональной структуры и структуры данных для данной предметной области. Кроме этого, анализ построенных знаний дает возможность выделить из них закономерности их построения, что позволяет создавать метазнания, которые будут хранить обобщенные знания, т.е. знания о знаниях, об особенностях структуры предметной области.

Использование описанного подхода в значительной степени сократит затраты на построение и оптимизацию как функциональной структуры предприятия, так и структуры данных, что позволит в существенной степени уменьшить длительность подготовительного этапа при построении моделей деятельности предприятия и снизить затраты на его проведение.

Применение изложенного метода позволит в значительной степени сократить сроки проектирования информационных систем, что особенно актуально для больших предприятий, внедряющих корпоративные информационные системы. Сокращение сроков проектирования даст возможность предприятию сэкономить средства при внедрении информационных систем. Решается также задача оптимизации процесса выделения бизнес-процессов при проведении реинжиниринга. Использование данного метода позволяет в значительной степени ускорить анализ предметной области с целью построения бизнес-процессов предприятия, который является одним из первых этапов на пути модернизации организации, улучшения управляемости и повышения эффективности ее деятельности.

Литература

1. Золотарёв О.В. Методы и инструменты моделирования предметной области // Цивилизация знаний: проблемы социальных коммуникаций: труды Междунар. науч.-практ. конф. – М. : РосНОУ, 2012.

2. Золотарёв О.В. Технология внедрения КИС: методические указания к лабораторным работам. – М. : РосНОУ, 2012.

3. Золотарёв О.В. Формализация знаний о предметной области на основе анализа естественно-языковых структур // Цивилизация знаний: проблема человека в науке XXI века». – М. : РосНОУ, 2012.

4. Золотарёв О.В. Управление в проектах внедрения распределенных корпоративных информационных систем // Вестник Российского нового университета. – 2012. – Выпуск 4.

5. Золотарёв О.В. Использование информационных технологий в реинжиниринге бизнес-процессов : методические указания к лабораторным работам. – М. : РосНОУ, 2013.

6. Золотарёв О.В. Инновационные решения в формировании функциональной структуры предметной области // Вестник РОСНОУ. – 2014. – Выпуск 4.

7. Шарнин М.М., Петров А.В., Кузнецов И.П. Методика учета интересов пользователя при работе в сети Internet на основе его профиля и ассоциативных связей.

8. Шарнин М.М., Сомин Н.В., Кузнецов И.П., Морозова Ю.И., Галина И.В. Автоматическое формирование ассоциативных портретов предметных областей на основе естественно-языковых текстов больших объемов для систем извлечения знаний.

9. Козеренко Е.Б. Стратегии выравнивания параллельных текстов: семантические аспекты // Информатика и её применения. – 2013. – Т. 7. – № 1. – С. 82–89.

10. Kozerenko, E.B. A Hybrid Model for Language Structures Disambiguation in Machine Translation. Proceedings of MLMTA'08. – Las Vegas: CRSEA Press, 2008.

11. Kuznetsov, I.P., Kozerenko, E.B. Linguistic Processor "Semantix" for Knowledge Extraction

from Natural Texts in Russian and English. Proceedings of MLMTA'08. – Las Vegas: CRSEA Press, 2008.

12. Kozerenko, E. Functional and Cognitive Aspects in Linguistic Modelling // Proceedings of ICAI'13, WORLDCOMP'13, July 22–25, 2013. – Las Vegas, Nevada, USA: CRSEA Press, USA.– 2013. – Vol. II. – P. 896–902.

13. Kuznetsov, I.P., Kozerenko, E.B., Somin, N.V. Semantic processor for knowledge extraction from texts in Russian and English // Proceedings of ICAI'13, WORLDCOMP'13, July 22–25, 2013. – Las Vegas, Nevada, USA: CRSEA Press, USA, 2013. – Vol. II. – P. 751–757.

14. Charnine, M., Petrov, A., Kuznetsov, I. Association-Based Identification of Internet User Interests // Proceedings of the 2013 International Conference on Artificial Intelligence (ICAI 2013), July 22–25, 2013. – Las Vegas, Nevada, USA: CSREA Press, 2013. – Vol. II. – P. 77–81.