УДК 004.41:47; 347.77

А.Ю. Дорошенко, Е.А. Оробинская, О.И. Король

ПОСТРОЕНИЕ ОНТОЛОГИЙ И ФРЕЙМВОРК ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ СОЗДАНИЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

Введение. Вызов сегодняшнего дня, обусловленный бурным ростом текстовых хранилищ в глобальной сети Интернет и технологическими проблемами их автоматической обработки, заключается в том, чтобы научить сами информационные системы (ИС) обнаруживать и правильно интерпретировать полезную информацию, предоставляемую текстом. В процессе работы над созданием современных компьютерных систем, решающих интеллектуальные задачи (в частности, понимание текстов на естественном языке), на первый план выдвигается проблема представления и извлечения знаний. Анализ любой текстовой информации, а особенно малоструктурированной, такой как патентно-конъюнктурная, и извлечение из полнотекстовых документов релевантных данных является актуальной задачей инженерии знаний в целом и онтологического инжиниринга в частности. Решение задачи обеспечения пользователей релевантной информацией в системе поиска и обработки определяется в основном правильным подбором инструментов делового анализа. Но немаловажным является и выбор инструментов поддержки процессов извлечения, преобразования, загрузки и хранения данных.

Цель работы – разработка лингвистического обеспечения для создания интеллектуальной системы автоматизированного построения онтологии на основе анализа и обработки текстовой ПКИ.

Качественное расширение возможностей ИС возможно при условии внедрения в них модулей, способных извлекать характеристики концептов на основе лингвистического анализа, что необходимо для более комплексной обработки.

Тенденция современных исследований направлена именно на внедрение в ИС лексических знаний. Например, для поиска таксономических (родо-видовых отношений) активно используются конструкции типа «Х, такие как У, У и У», где Х – понятие более высокого уровня (гипероним), У – уточняющее понятие (гипоним). Ряд работ, как например [4], посвящено определению конкретных лингвистических шаблонов (patterns), специфичных для некоторой ПО и обеспечивающих качественное обнаружение в тексте релевантных данных.

В общепринятом смысле под системой понимается множество взаимосвязанных элементов, обособленное от среды и взаимодействующее с ней, как целое. Несложно видеть, что патентная конъюнктурная информация, как и любая другая прикладная область, действительно представляет собой специальное множество с эмерджентными свойствами, обладающее структурной, функциональной и динамической организацией [5, 6].

Однако, сначала следует доказать справедливость утверждения, о том что языковые феномены (тексты в том числе), могут (и должны) быть описаны с позиций естествознания как системные.

1. Иллюстрация работы лингвистического модуля

Ниже приводится описание фрагмента работы лингвистического модуля для онтологии ПКИ.

Данный модуль распознает:

I. Таксономические отношения типа is-a, выраженные формулой “такие Х, как У,У,…, У и/или У”, где Х – это гипероним, уточняемый рядом гипонимов У; а также все ее возможные синтаксические формы. Следствие из этого правила: все гипонимы одного гиперонима являются синонимами.

II. Таксономические отношения типа is-a, выраженные формулами «У является Х (в Т.п)» или «У представляет [собой] Х (в В.п.)», ”, где Х – это гипероним для У (гипоним).

III. Отношения эквивалентности для одинаковых частей речи, идущих подряд и связанных союзами и, или, также, а также;

IV. Свойства концептов на основании того, что качество субъекта (квалитатив) в русском языке может выражаться такими формулами:

1. вспомогательными глаголами {отличаться, выражаться, проявляться, характеризоваться, обладать, снабжаться} + Т.п. признакового слова (так называемый предицирующий компонент);
2. с помощью родительного падежа, когда предицирующий компонент определен сочетанием {прилагательное + существительное};
3. с помощью связанной синтаксемы с родительным падежом, обозначающей объект при отглагольном существительном;
4. в эксплицитной форме, конструкцией типа «объект может иметь такие {характеристики, свойства, признаки}.

ИС снабжена парсером русского языка АОТ [3], способным распознавать грамматические формы встречаемых слов.

Для правильной работы необходимо также определить порядок применения (приоритет) правил.

В данном случае порядок будет следующий:

1. Определяются синонимичные конструкции (правило III).
2. Проверяется роль Р.п. падежа в порядке убывания приоритета IV.2˜ IV.3.
3. Обнаруживаются таксономические конструкции (правило I).
4. Обнаруживаются таксономические конструкции (правило II).
5. Обнаруживаются свойства концептов по правилу IV.1 или IV.4.

На вход ИС подается для анализа следующий текст:

*«Предлагаем компьютерный **стол**, который представляет собой мебельное изделие жесткой конструкции в виде тумбы. **Стол** содержит коробчатый корпус, а также ящики, установленные в корпусе на направляющих друг над другом. По крайней мере, два соседних ящика установлены на телескопических направляющих. В корпусе **стола** предусмотрено отверстие для прохода кабелей. Нижний из двух ящиков **стола**, установленных на телескопических направляющих, смещен книзу, а его передняя стенка удлинена на такое же расстояние вверх, до уровня донной части верхнего ящика»*.

В качестве входного термина, чьи свойства мы пытаемся обнаружить, используем слово «стол». Т.е. система будет применять описанные выше правила только к тем предложениям, в которых обнаружит слово «стол».

Далее приведены отдельные свойства термина «стол», которые обнаруживаются в тексте информационной системой:

«*компьютерный стол*» - правило IV.2; после нормализации получим: термин: стол; имеет свойство: компьютерный;

«*…изделие жесткой конструкции*»˜ правило IV.3; после нормализации получим: термин: изделие; имеет свойство: жесткая конструкция;

«*…стол … представляет собой изделие…*»˜ правило II; термин: стол является разновидностью: изделие (стол˜ is-a˜ изделие);

«*…коробчатый корпус, а также ящики…*»˜ правило III; после нормализации получим ряд однотипных элементов: корпус, ящик.

Отметим, что важным преимуществом предлагаемого метода является то, что «точками входа» для анализа текста могут служить и слова, и собственно синтаксемы.

2. Фреймворк для построения онтологии ПКИ

Авторы предлагают новый фреймворк ИС (рис.1), обеспечивающий автоматическое извлечение предопределенных свойств концептов и построение онтологии ПКИ, сочетающий в себе быстроту статистических методов и точность лингвистического подхода с позиций синтаксем.

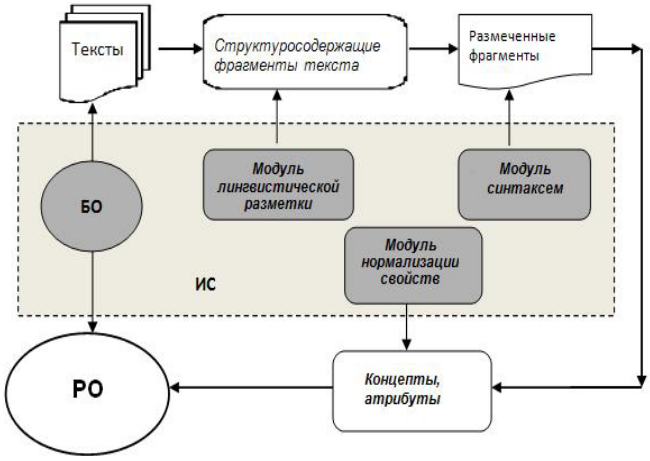

Рис.1. Фреймворк ИС на основе синтаксем

Разрабатываемая ИС состоит из следующих модулей:

• Модуль базовой онтологии (БО), представляющий собой простую таксономию основных концептов ПО. БО одновременно служит «фильтром» для отбора фрагментов текста, потенциально содержащих информацию, которая позволит либо добавить новые свойства для существующих концептов БО, либо добавить новые концепты в БО (с их свойствами) и получить расширенную онтологию (РО). подмодуля концептов и их иерархических отношений;

• Модуль лингвистической разметки, основой которого служит парсер языка. На выходе этого модуля получаем фрагменты текста, снабженные необходимой синтаксической информацией для дальнейшей обработки;

• Модуль синтаксем русского языка (основной модуль ИС), включающий в себя словарь типа WordNet (аналоги для русского языка RuNet, RussNet), объединяющий слова в группы синонимов;

• Модуль нормализации свойств, преобразующий найденные на предыдущем этапе данные в формальное представление на OWL и обеспечивающий расширение БО.

Задача обнаружения терминов-кандидатов довольно успешно решается сегодня многочисленными статистическими методами [2]. Задачи их группировки также отчасти разрешимы этими методами (на основе анализа частот совместного появления слов на некотором ограниченном расстоянии). Но проблема обнаружения связей между понятиями не может быть удовлетворительно решена без привлечения лингвистических знаний.

Для обеспечения языковой компетентности достаточной для самообучения и решения конечной задачи, т.е. построения патентной онтологии на базе текста, ИС сама должна обладать знаниями соответствующего порядка общими (языковыми) и специальными, (относящимися к нашей предметной области). Такая ИС должна, по сути, объединять в себе две онтологии: общую онтологию языка (русского) и базовую (стартовую) онтологию ПКИ.

Хотя все еще не существует полной онтологии какого-либо естественного языка, однако, разработки программного обеспечения для задач онтологического инжиниринга демонстрируют успехи в именно этом направлении. Так многочисленные электронные словари-тезаурусы (MRD – machine readable dictionary), с лексикографической разметкой (какие как WordNet и его русскоязычные аналоги RussNet, RuNet и др.) можно рассматривать как элементы C и $\leq$c требуемой лингвистической онтологии. А парсеры языка (проект АОТ) могут быть успешно использованы как для определения ролей концептов, так и для определения их атрибутов. То чего действительно не хватает для полноценной интеллектуальной системы, так это систематизированного репертуара лексико-синтаксических единиц языка, несущих в себе однозначно трактуемую семантику и одновременно выполняющих роль «элементарных единиц сборки» высказываний (текстов). Пересечение такого репертуара «архе-функций», несомых синтаксической формой и репертуара слов-носителей категориальных референций, дает проекцию однозначно трактуемой роли (функции) выполняемой данным понятием.

Теперь приступим к формальному описанию предложения. Для того, чтобы увидеть в предложении формулу алгебры предикатных операций, сначала представим в виде граф-схемы синтаксическую структуру какой-нибудь формулы [1].

Возьмем, к примеру, формулу алгебры булевых функций $\overline{X}_1 X_2 \vee X_3 \overline{X}_4$. Ее можно выразить графически схемой, изображенной на рис. 2. Кружки со знаками булевых операций $\neg$, $\wedge$ и $\vee$ изображают преобразователи формул.

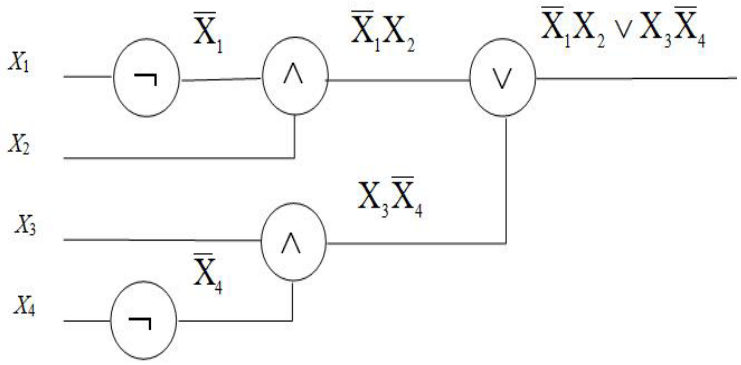

Рис. 2. Графическое представление формулы алгебры булевых функций

Схема синтезирует формулу $\overline{X}_1 X_2 \vee X_3 \overline{X}_4$ из ее аргументов X_1, X_2 X_3, X_4. Так, проходя через крайний справа блок дизъюнкции, формулы $\overline{X}_1 X_2$ и $X_3 \overline{X}_4$ преобразуются в формулу $\overline{X}_1 X_2 \vee X_3 \overline{X}_4$. Та часть формулы, на которую бинарная операция ($\wedge$ или $\vee$) действует первой, поступает на преобразующий блок по горизонтальному входу, второй – по вертикальному. Схема формулы представляет собой древовидный граф [7].

Приведем пример логической сети для синтаксического разбора предложения в полнотекстовых документов. Логическая сеть строится для конкретного типа предложения. Например, для предложения «Висока частота поширеності захворювань потребує удосконалення знань у цій галузі» структура логической сети представлена на рис. 3.

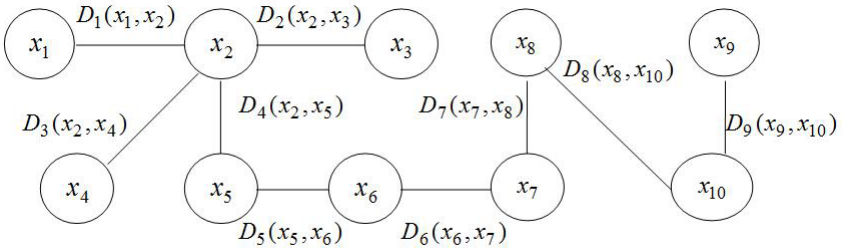

Рис. 3. Структура логической сети

Предметными переменные для данного предложения являются сами слова этого предложения, т.е. $x_1 - висока$, $x_2 - частота$, $x_3 - поширеності$, $x_4 - захворювань$, $x_5 - потребує$, $x_6 - удосконалення$, $x_7 - знань$, $x_8 - у$, $x_9 - цій$, $x_{10} - галузі$.

Область изменения переменных формируется экспертом (в нашем случае он является одновременно и пользователем):

$A(x_1) = \{висока, низька, середня\}$,

$B(x_2) = \{частота, амплітуда, швидкість\}$,

$C(x_3) = \{поширеність\}$,

$D(x_4) = \{захворювання, віруси, ОРВІ\}$,

$E(x_5) = \{потребувати, вимагати, спостерігати\}$,

$F(x_6) = \{удосконалення, поліпшення\}$,

$G(x_7) = \{знання, уміння, досвід\}$,

$H(x_8) = \{у, до, на, від\}$, $R(x_9) = \{ця, сучасна, актуальна\}$, $S(x_{10}) = \{галузь, сфера, проблема\}$.

Бинарные связи между узлами логической сети описываются уравнениями следующего вида:

$$висока\_частота(x_1,x_2) = висока(x_1) \wedge частота(x_2) \wedge D_1(x_1,x_2),$$

$$частота\_поширеності(x_2,x_3) = частота(x_2) \wedge поширеність(x_3) \wedge D_2(x_2,x_3),$$

$$частота\_захворювань(x_2,x_4) = частота(x_2) = частота(x_2) \wedge захворювання(x_4) \wedge D_3(x_2,x_4),$$

$$частота\_потребує(x_2,x_5) = частота(x_2) \wedge потребувати(x_5) \wedge D_4(x_2,x_5),$$

$$потребує\_удосконалення(x_5,x_6) = потребувати(x_5) \wedge удосканалення(x_6) \wedge D_5(x_5,x_6),$$

$$удосконалення\_знань(x_6,x_7) = удосконалення(x_6) \wedge знання(x_7) \wedge D_6(x_6,x_7),$$

$$знань\_у(x_7,x_8) = знання(x_7) \wedge у(x_8) \wedge D_7(x_7,x_8),$$

$$у\_галузі(x_8,x_{10}) = у(x_8) \wedge галузь(x_{10}) \wedge D_8(x_8,x_{10}),$$

$$цій\_галузі(x_9,x_{10}) = ця(x_9) \wedge галузь(x_{10}) \wedge D_9(x_9,x_{10}).$$

Установка начальных состояний полюсов считается нулевым тактом. Источниками на первом такте будут все те полюса, в которые внес информацию пользователь на нулевом такте. Каждый такт включает в себя выполнение следующих действий.

Для каждого источника вычисляются все исходящие функции, которые с ним связаны. Аргумент этих функций – состояние данного полюса.

Для каждого приёмника вычисляется новое состояние: образуется пересечение старого состояния и значений всех активных входящих функций.

Для каждого приёмника сравниваются предыдущее и новое состояния: те из приёмников, для которых состояния не совпали, будут источниками на следующем такте. Если множество источников для следующего такта оказалось пустым (т.е. совпадение имеет место для всех приёмников), то работа сети прекращается – решение найдено.

Решением являются состояния всех полюсов сети после завершения работы алгоритма. Таким образом, алгоритм логической сети можно отнести к семейству волновых алгоритмов.

Заключение. 1. Обоснована возможность рассматривать текст как динамическую систему и, следовательно, применять естественнонаучные методы для его моделирования. 2. Описан ключевой подход к семантическому анализу на основе синтаксем. 3. Разработана логическая сеть для синтаксического анализа на базе моделей сочетаемости слов, которая опирается на структуру предложения и семантику текста в целом, которое разрешило перейти к единой математической модели обработки текстовой информации в документах с использованием единого математического аппарата. 4. Показано использование алгебры конечных предикатов для решения задачи автоматического моделирования ПКАИ. Приведены графические представления формул алгебры булевых функций. На примерах, показаны схемы формул предложений.

ЛИТЕРАТУРА:

1. Оробинская Е.А. Языковая компетенция информационных систем / Е.А. Оробинская, О.И. Король, Н.В. Шаронова // Вісник Національного технічного університету "Харківський політехнічний інститут". Проблеми інформатики і моделювання. – Х.: НТУ «ХПІ», – 2012.
2. Федотов Н. Н. Средства информационного обеспечения автоматизированных систем управления / Н. Н. Федотов, Л. Б. Венчковский. – М. : Изд-во стандартов, 1989. – 192 с.
3. Бондаренко М.Ф. Теория интеллекта: учеб./ М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. Харьков: Компания СМИТ, 2006. - 576 с.
4. Шаронова Н.В. Автоматизированные информационные библиотечные системы: задачи обработки информации: монография, НУА / Н.В. Шаронова, Н.Ф. Хайрова; – Харьков 2003, 120 с.
5. Зубов А. В. Основы искусственного интеллекта для лингвистов / А. В. Зубов, И. И. Зубова. – М. : Университетская книга; Логос, 2007. – 320 с.
6. Ермаков А.Е. Автоматизация онтологического инжиринга в системах извлечения знаний из текста /А. Е. Ермаков // труды Международной конференции Диалог'2008. – Москва, Наука, 2008. - С. 136-140.
7. Канищева О. В. Использование алгебры предикатных операций для описания естественно-языковых отношений / О. В. Канищева // Інформаційні технології: наука, техніка, технологія, освіта, здоров'я : матеріли XVII міжнар. наук.-практ. конф. – Харків : НТУ «ХПІ», 2009. – С. 16.

ДОРОШЕНКО Анастасия Юрьевна – аспирант Национального технического университета «Харьковский политехнический институт», кафедра интеллектуальных компьютерных систем.

Научные интересы: прогрессивные информационные технологии.

ОРОБИНСКАЯ Елена Александровна – аспирант Национального технического университета «Харьковский политехнический институт», и Университета им. Люмьер Лион-2 (Лион, Франция), кафедра интеллектуальных компьютерных систем.

Научные интересы: прогрессивные информационные технологии.

КОРОЛЬ Ольга Игоревна – аспирант Национального технического университета «Харьковский политехнический институт» и Университета им. Люмьер Лион-2 (Лион, Франция), кафедра интеллектуальных компьютерных систем.

Научные интересы: прогрессивные информационные технологии.