УДК 004.912

Л.Э. ЧАЛАЯ, А.В. ЧИЖЕВСКИЙ

МЕТОД АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ С ДРЕВОВИДНОЙ СТРУКТУРОЙ КОНЦЕПТОВ

Предлагается метод автоматического построения онтологической модели данных на основе анализа корпуса научных текстов для определенной предметной области, что позволяет формировать онтологию с древовидной структурой концептов с учетом семантических связей между ними. Осуществляется программная реализация метода, предусматривающая возможность последовательного построения древовидной онтологической модели по исходному корпусу монотематических текстов.

1. Постановка задачи

Онтологии предметной области в настоящее время широко применяются в области построения поисковых систем, систем представления знаний, инженерии знаний и при решении задач семантической интеграции информационных ресурсов. Под онтологией понимается «формальная спецификация концептуализации, которая имеет место в некотором контексте предметной области» [1]. В свою очередь, концептуализация представляет собой представление предметной области через описание множества понятий (концептов) предметной области и связей (отношений) между ними. В последние годы сформировалась парадигма компьютерных онтологий, основными признаками которых являются [2]:

– иерархическая структура конечного множества понятий, описывающих заданную предметную область;

– структура онтологии может быть представлена онтографом, вершинами которого являются понятия (концепты), а дугами – семантические отношения (связи) между ними;

– концепты и связи между ними интерпретируются в соответствии с результатами анализа электронных источников знаний (например, корпуса текстов) заданной предметной области;

– онтограф должен быть представлен формально на одном из языков описания онтологий.

В общем случае компьютерная онтология некоторой предметной области формально представляется упорядоченной тройкой [3]:

$$O =< X, R, F >, \quad (1)$$

где X, R, F – конечные множества соответственно: X – концепты (понятия, термины) предметной области; R – отношения между ними;, F – функции интерпретации X и/или R.

Рассмотрение граничных случаев множеств в (1): $R = \varnothing$; $R \neq \varnothing$; $F = \varnothing$; $F \neq \varnothing$ во всех четырех комбинациях значений R и F дает различные варианты онтологических конструкций, начиная от простого словаря и таксономии до формальной структуры концептуальной базы знаний для высокоинтеллектуальных знание-ориентированных систем [4].

По своей функциональной полноте и степени формальности различают три вида онтологий: простая, полная (или строгая) и множество промежуточных или неполных онтологий.

Простая – это такая онтология, в которой $R = \varnothing$; $F = \varnothing$. Она служит (в основном) для однозначного восприятия научным сообществом понятий в соответствующей прикладной области.

Строгая или полная ($R \neq \varnothing, F \neq \varnothing$) – это такая онтология, в которой множества концептов и концептуальных отношений являются максимально полными, а к функциям интерпретации добавляются аксиомы, определения и ограничения. При этом описания всех компонент представлены на некотором формальном языке, доступном для их интерпретации компьютером. Схема формальной модели полной онтологии описывается четверкой:

$$O =< X, R, F, A(D, R_s) >, \quad (2)$$

где X – множество концептов; R – множество концептуальных отношений между ними; $F : X \times R$ – конечное множество функций интерпретации, заданных на концептах и/или отношениях; A – конечное множество аксиом, которые используются для записи всегда истинных высказываний (определений и ограничений); D – множество дополнительных определений понятий; R_s – множество ограничений, определяющих область действия понятийных структур.

В промежуточных или неполных онтологиях ($R = \varnothing$, $F \neq \varnothing$; $R \neq \varnothing$, $F = \varnothing$) для каждого концепта (или их большей части) добавлены аксиомы и определения, представленные на естественном языке (ЕЯ). Одним из распространенных вариантов неполной онтологии является структура $O =< X, R >$, где множество F в явном виде отсутствует ($F = \varnothing$), в предположении, что концепты общеизвестны (определены по умолчанию) либо (и) достаточно полно интерпретированы отношениями R .

Применение онтологического подхода для автоматической обработки текстов на естественном языке предполагает сопоставление понятиям онтологии предметной области (к которой принадлежит множество текстов) языковых выражений (слов и словосочетаний), которыми понятия могут быть выражены в текстах. При этом структура концептов такой онтологии, представленных словосочетаниями из корпуса текстов, является древовидной, причем конкретный вид дерева определяется сложностью словосочетания.

Качество формируемых онтологий, используемых для создания поисковых систем, во многом определяется полнотой учета в онтологической модели наиболее значимых концептов для корпуса анализируемых текстов с учетом их тематической специфики (под концептами будем в дальнейшем понимать наиболее значимые слова и словосочетания в анализируемом тексте, которые могут быть учтены в онтологической модели). В связи с этим целесообразно решить задачу формирования множества концептов будущей онтологии типа $O =< X, R >$ с учетом связей между ними (при возможности оценивания концептов по двум атрибутивным показателям, связанным с анализом исходного корпуса текстов предметной области: TF и TF/IDF) . Процедура определения шаблонных связей типа «часть-целое» была рассмотрена в работе [5]. Более сложной является задача нахождения связей типа «отношение» между выбранными концептами. В значительной степени именно задание таких отношений наделяет онтологию интеллектуальным смыслом, что позволяет компьютеру, работающему с такой онтологией, максимально эффективно использовать заложенные в нее знания. Следует отметить, что при поиске слов и словосочетаний, которые могут применяться в качестве концептов, сформированное множество концептов-претендентов не всегда соответствует такому же множеству, составленному экспертом предметной области. Это приводит к тому, что некоторые важные понятия предметной области могут не попасть в автоматически создаваемую онтологию. Кроме того, в этих методах отсутствует процедура общего ранжирования по значимости списка всех концептов-претендентов, а осуществляется лишь раздельное ранжирование слов и словосочетаний, входящих в этот список.

В частности, возникают следующие проблемы: не всегда удается правильно найти связи между концептами; не всегда удается выделить концепты, имеющие связь с наибольшим количеством других концептов; найденные связи между концептами будущей онтологии не всегда актуальны для конкретной предметной области. При этом не только повышается используемый объем памяти и увеличивается время на создание онтологии и обработку запросов к ней, но и избыточным становится объем онтологии, что снижает оперативность дальнейшего ее применения. Целесообразно рассмотреть возможность устранения перечисленных трудностей на основе комбинированного применения и модификации существующих методов определения релевантных связей между концептами формируемых онтологических моделей.

Целью данного исследования является разработка и программная реализация метода автоматического построения онтологий с древовидной структурой концептов-словосочетаний для заданной предметной области по корпусу текстов.

2. Метод построения онтологии

Концепция предлагаемого метода автоматического поиска древовидной структуры концептов формируемой онтологии состоит в реализации следующих этапов:

Этап 1. Выделение концептов высшего уровня онтологического дерева.

Этап 2. Выделение ключевых слов в концептах высшего уровня.

Этап 3. Установление связей между концептами высшего уровня.

Этап 4. Определение дочерних концептов.

Этап 5. Установление связей между концептами онтологии.

Этап 6. Построение онтографа.

Очевидно, что количество дочерних концептов для каждого из концептов высшего уровня может быть различным и общая структура онтологического дерева будет зависеть от предметной области и репрезентативности исходного корпуса текстов.

Рассмотрим подробнее этапы предлагаемого метода.

Выделение концептов высшего уровня онтологического дерева. Выделение концептов высшего уровня (КВУ) для любой онтологии предметной области (ОПО) является важным этапом в общем алгоритме проектирования, так как древовидные ОПО строятся фрагментарно, начиная от исходной вершины, и именно качество задания КВУ определяет в дальнейшем эффективность объединения концептов и связей вершин в общую онтологию. Очевидно, что концепты низших (дочерних) уровней ОПО являются подклассами КВУ и поэтому наследуют признаки понятия-класса, если они связаны между собой отношениями частичного порядка.

Под концептами высшего уровня формируемой онтологии будем понимать наиболее важные (значимые) для рассматриваемой предметной области словосочетания W_i, $i = H$ (H – число концептов высшего уровня) из исходного корпуса текстов. Эти словосочетания являются элементами концептуальной составляющей формируемой онтологии вида $O =< X, R >$, где $X = W$.

В работе [6] был предложен метод поиска таких словосочетаний, основанный на количественной оценке важности элементов текста с использованием значений TF, TF/IDF и так называемых рангов слов. Исходной операцией здесь является предварительное упорядочение по убыванию важности и составление соответствующих списков важных слов для показателей TF и TF/IDF. Под частными рангами R_1 и R_2 некоторого слова понимаются значения величин, обратных номерам позиции этого слова в упорядоченных списках для TF и TF/IDF соответственно. Под общим рангом R некоторого слова будем понимать коэффициент, который соответствует наибольшему из значений частных рангов (R_1 и R_2) этого слова для анализируемого корпуса текстов.

Предлагается оценивать в анализируемых текстах значимость слов w_i в i-м тексте корпуса по значениям коэффициента $K(w_i)$, рассчитываемого по следующей зависимости:

$$K(w_i) = \frac{1}{|R_1 - R_2|} * \frac{Q(w_i)}{N_{max} - N_{min}}, \qquad (3)$$

где $Q(w_i)$ – количество текстов, в которых содержится слово w_i; N_{max}, N_{min} – наибольшее и наименьшее число вхождений слов w_i в корпус текстов соответственно.

Согласно (3) значимые термины, выделяемые для создания отношений между словами-концептами, должны встречаться в большинстве текстов корпуса и при этом быть максимально равномерно распределены в каждом тексте корпуса текстов.

После определения значений R_1 и R_2 производим модификацию исходных списков. При этом слово относим к первому модифицированному списку, если для него $R_1 > R_2$, и, соответственно, ко второму, если $R_2 > R_1$.

Очевидно, что важность словосочетаний для текущего текста зависит от наличия в них слов каждого из модифицированных списков. Например, словосочетание, полностью составленное из слов первого списка, недостаточно хорошо отображает семантику текста,

так как все слова в нем, скорее всего, окажутся слишком общими. Однако, если словосочетание составлено полностью из слов второго списка, в нем могут отсутствовать ключевые слова, имеющие максимальную частоту вхождения в документы анализируемого корпуса текстов.

Важно также правильно выбрать максимальную длину оцениваемых словосочетаний. Очевидно, что с ее увеличением уменьшается вероятность присутствия в тексте осмысленных словосочетаний заданной длины.

В связи с этим представляется целесообразным в общий критерий количественной оценки важности словосочетаний ввести нормированный коэффициент $K(W_i)$, который будет зависеть как от количества вхождений в словосочетание w_i слов из первого и второго модифицированных списков, так и от длины словосочетания.

Анализ списков словосочетаний (для представительного набора текстов по направлению «Компьютерные науки») позволил экспериментально оценить влияние длины словосочетаний, а также количества слов из первого и второго модифицированных списков на вероятность присутствия этих словосочетаний в документах анализируемого корпуса текстов. Очевидно, что такая вероятность может быть принята в качестве коэффициента $K(W_i)$, значения которого находятся в диапазоне [0; 1].

В таблице представлены значения коэффициента $K(W_i)$ для разных типов словосочетаний. Предлагается оценивать важность словосочетания, перемножая ранги отдельно взятых слов. При этом, поскольку ранги слова нормированы от нуля до единицы, то с увеличением длины словосочетания уменьшается результат такого произведения. В связи с этим его необходимо умножить на количество слов в рассматриваемом словосочетании.

Следует также учитывать, что слова, для которых $R_1 > R_2$, определяют специфику анализируемого текста, т.е. слова из первого списка с большой вероятностью попадут в большинство словосочетаний, характерных для этого текста. Следовательно, оценки важности таких словосочетаний будут близки. Наиболее же специфичными для конкретного текста являются слова второго списка, для которых $R_2 > R_1$, следовательно, именно за счет этого можно повысить релевантность критерия оценки важности словосочетания.

Значения коэффициента $K(W_i)$ для разных типов словосочетаний

№	Количество слов в словосочетании по первому списку	Количество слов в словосочетании по второму списку	$K(W_i)$
1.	2	0	0,3
2.	1	1	1
3.	3	0	0,2
4.	2	1	0,6
5.	1	2	0,7
6.	3	1	0,5
7.	1	3	0,8

Кроме того, величина модуля разности рангов слов по разным спискам ($|R_{1j} - R_{2j}|$) влияет на вероятность того, что слово, которому соответствует максимальный ранг, является важным для анализируемого корпуса текстов: чем больше минимальный модуль разности двух рангов одного и того же слова, тем выше находятся все слова из словосочетания в каком-либо из двух списков и тем важнее словосочетание в целом.

Для комплексной оценки важности словосочетания W_i в рассматриваемом тексте можно использовать следующий коэффициент $M(W_i)$:

$$M(W_i) = K(W_i) * (\prod_{i=1}^{n} R_i) * \min(|R_{1j} - R_{2j}|) * n * \max(TF / IDF_j)\ ;\ j = 1,..,n\ ,\ R_{1j} \neq R_{2j}\ , \qquad (4)$$

где n – количество слов в словосочетании W_i (не считая стоп-слов); TF/IDF – коэффициенты оценки важности слова; R_i, R_{1j}, R_{2j} – общий и частные ранги слова w_j из словосочетания W_i соответственно.

Важность словосочетаний $M(W_i)$ предлагается оценивать по произведению рангов соответствующих слов. При этом, поскольку ранги слова нормированы от нуля до единицы, с увеличением длины словосочетания уменьшается результат такого произведения. В связи с этим его необходимо умножить на количество слов в рассматриваемом словосочетании. Следует также учитывать, что слова, для которых $R_1 > R_2$, определяют специфику анализируемого текста, т.е. слова из первого списка с большой вероятностью попадут в большинство словосочетаний, характерных для этого текста. Наиболее же специфичными для конкретного текста являются слова второго списка, для которых $R_2 > R_1$, следовательно, именно за счет этого можно повысить релевантность критерия оценки важности словосочетания.

Общий алгоритм поиска наиболее важных словосочетаний (для заданного корпуса текстов) состоит в следующем:

Шаг 1. Предварительно ранжируются по важности исходные списки слов по заданному корпусу электронных текстов (по R_1 и R_2) и определяются их общие ранги.

Шаг 2. Производится модификация ранжированных списков по приведенным выше правилам.

Шаг 3. Формируется исходный список словосочетаний из слов модифицированных списков.

Шаг 4. Определяется оценка важности словосочетаний $M(W_i)$.

Таким образом, в концептах высшего уровня могут быть использованы словосочетания с наибольшей оценкой $M(W_i)$ либо отдельные слова из этих словосочетаний, соответственно с большей оценкой важности слова $K(W_i)$.

Выделение ключевых слов в найденных концептах. В концептах высшего уровня могут быть выделены главные смысловые слова (зачастую это термины предметной области) и словосочетания, определяемые в соответствии с результатами ранжирования связанных концептов (по методу главного компонента). Для анализа смысла конструкции тройки «концепт – связь – концепт» в алгоритме формирования онтологии важно привести рассматриваемое словосочетание к одному слову, что позволит представить концепт максимально абстрактно и определить, является ли существенной для будущей онтологии связка с рассматриваемым глаголом. Конструкцию словосочетания в общем случае можно представить как главное слово и зависимые члены. С помощью синтаксического анализатора несложно выделить в словосочетании такие элементы как существительное, числительное или местоимение (при наличии нескольких идентичных элементов в качестве главного принимается слово, приведенное в именительном падеже).

Определение дочерних концептов и установление связей между ними. Следующим шагом создания дерева иерархии будущей онтологии является нахождение дочерних концептов по ключевым словам высшего уровня. В качестве дочерних концептов словосочетаний выделяются словосочетания, имеющие наибольшее совпадение слов с ключевыми словами высшего уровня и, соответственно, наибольшую оценку $M(W_i)$. На следующем этапе выявляются связи между найденными концептами (по методу главного компонента). Отметим, что для каждого из полученных дочерних концептов могут существовать подчиненные дочерние концепты более низкого уровня (количество уровней обусловливается спецификой предметной области и репрезентативностью исходного корпуса текстов). Алгоритм их выделения в целом аналогичен рассмотренному выше алгоритму поиска концептов высшего уровня, но при этом связан с необходимостью учета значимых связей между концептами смежных уровней.

Установление связей между концептами онтологии. Выделим три основных подхода для решения задачи установления связей между концептами проектируемой онтологии:

– поиск слова-претендента на связь в онтологии и последующий подбор концептов, для которых актуальна эта связь (метод 1);

– определение для рассматриваемого концепта списка вероятных слов-претендентов на использование в качестве связи для этого концепта и последующий подбор концепта для установления связи (метод 2);

– нахождение в онтологии двух концептов, которые необходимо связать, и последующий подбор связи для данных концептов (метод 3).

Достоинства первого подхода (метод 1):

– поиск в тексте слов-связей и концептов осуществляется раздельно. Это означает, что концепт и связь не обязательно должны составлять в тексте словосочетание при поиске данной связки в тексте программы автоматического синтеза онтологии;

– возможность варьировать количество учитываемых связок «концепт-связь-концепт» с помощью настраиваемых коэффициентов (уменьшать в случае нахождения большого количества ненужной информации и увеличивать при недостаточном количестве связей в онтологии).

Недостатки первого подхода:

– в общем множестве найденных связей между концептами присутствуют несущественные или несуществующие связи;

– некоторые важные концепты предметной области не имеют связей сформированного множества с другими концептами проектируемой онтологии.

Устранению отмеченных недостатков способствует комбинированное применение второго и третьего подходов (методы 2 и 3).

В работе [6] был предложен метод поиска связей для онтологии, основанный на таком комбинированном подходе, названный методом главного концепта.

Этот метод предполагает необходимость вычисления вероятности применения слова в качестве релевантной связки для рассматриваемого концепта. В качестве слов-связок могут применяться как слова, специфичные для рассматриваемой предметной области, так и достаточно общие слова, которые могут присутствовать в любом тексте. Можно отметить, что слово-связка вероятнее всего будет находиться в тексте между понятиями, которые оно связывает. Вследствие этого целесообразно определить степень специфичности претендента на слово-связку в контексте понятия, которое будет связывать данное слово-связка. В соответствии с предлагаемым алгоритмомсчитают, что если слово-связка входит в контекст леммы-понятия в рамках рассматриваемого текста, то оно специфично в контексте данного понятия. Введем понятие тройки элементов, используемых для реализации процедуры предварительного отбора наиболее релевантных связок для проектируемой онтологии. К элементам такой тройки отнесем: слово (словосочетание), обозначающее связь между двумя концептами (L_1), и собственно два концепта (W_1 и W_2), каждый из которых может быть представлен одним словом либо словосочетанием. Таким образом, тройку можно представить в виде: «слово№1, связь, слово№2»:

$$W_1 \leftrightarrow L_1 \leftrightarrow W_2 . \quad (5)$$

Отметим, что если концепт представлен словосочетанием, то в тройку вносится главное слово словосочетания.

Выделим четыре возможных варианта представления любой тройки в зависимости от уровня специфичности слова-связки по отношению к понятиям:

$$W_1 \xleftrightarrow{F_{(W1,L1)}\uparrow} L_1 \xleftrightarrow{F_{(W2,L1)}\uparrow} W_2 ; \quad (6)$$

$$W_1 \xleftrightarrow{F_{(W1,L1)}\uparrow} L_1 \xleftrightarrow{F_{(W2,L1)}\downarrow} W_2 ; \quad (7)$$

$$W_1 \xleftrightarrow{F_{(W1,L1)}\downarrow} L_1 \xleftrightarrow{F_{(W2,L1)}\uparrow} W_2 ; \quad (8)$$

$$W_1 \xleftrightarrow{F_{(W1,L1)}\downarrow} L_1 \xleftrightarrow{F_{(W2,L1)}\downarrow} W_2 , \quad (9)$$

где символ $F_{(W_i,L1)}\uparrow$ означает, что слово-связка L_1 специфично для концепта W_i, а $F_{(W_i,L1)}\downarrow$ означает, что слово-связка L_1 не специфично для концепта W_i.

На основе статистического анализа текстов рассматриваемой предметной области могут быть определены коэффициенты вероятности p_1, p_2, p_3, p_4 принадлежности определенной тройки к одному из вариантов ее представления: (6), (7), (8) или (9). На основании полученных значений p_1, p_2, p_3, p_4 определим вероятности выбора слова в качестве связки в зависимости от его положения в предложении по отношению к концептам.

При принятии решения о занесении той или иной тройки в проектирующуюся онтологию, кроме расположения элементов тройки, необходимо учитывать наличие слов между ними и их количество. Очевидно, что целесообразнее вносить в онтологию тройки, элементы которой следуют непосредственно друг за другом, чем тройки, между концептами и связкой которой находятся фрагменты предложения. Назовем расстоянием между элементами тройки количество слов, которые находятся в предложении между двумя любыми элементами тройки. Обозначим через N расстояние в предложении между двумя концептами W_1 и W_2 рассматриваемой тройки.

Тогда вероятность актуальности рассматриваемой тройки в зависимости от положения ее элементов в предложении можно определить следующим образом:

$$P_{place} = \frac{k * \left(\left| m - n \right| / \min(n, m) + 2) + 1 \right)}{n + m + 1}, \qquad (10)$$

где n – расстояние от L_1 до W_1, $n = N$, если между L_1 и W_1 находится W_2; m – расстояние от L_1 до W_2, $m = N$, если между L_1 и W_2 находится W_1.

В соответствии с (10), чем больше расстояние между словом-связкой и концептами в тройке, тем меньше вероятность ее актуальности для проектируемой онтологии. Также необходимо отметить, что приведенная формула учитывает приоритет троек, у которых расстояние слова-связки хотя бы с одним из концептов является намного меньше среднего значения такого расстояния для всей совокупности рассматриваемых концептов.

Алгоритм установления связей между концептами онтологии по методу главного концепта можно представить набором следующих действий:

– выбор концепта/понятия и нормализация его до одного слова (W_1), для которого следует сформировать тройку в проектируемой онтологии;

– определение множества слов $M\uparrow(W_1)$, входящих в контекстное множество данного концепта W_1 (из множества всех слов в предложениях, где присутствует данный концепт с понятием $M(W_1)$, а также множества слов $M\downarrow(W_1)$, не входящих в контекстное множество данного концепта;

– определение наиболее вероятного типа связи ($F_{(W_i,L_1)}\uparrow$ или $F_{(W_i,L_1)}\downarrow$) между концептом W_1 и предполагаемым словом-связкой L_1;

– определение множества $M(L_i)$, состоящего из претендентов на слова-связки, удовлетворяющих установленному типу связи ($M(L_1)$, принимается как $M\uparrow(W_1)$ или как $M\downarrow(W_1)$);

– определение для каждого L_i (из множества $M(L_i)$) множества слов $M\uparrow(W_2)$, входящих в контекстное множество данного слова L_i (из множества всех слов, входящих в одно предложение с данным словом L_i и данным словом W_1, во всех предложениях, где присутствуют L_i и W_1), а также множества слов $M\downarrow(W_2)$, не входящих в контекстное множество данного концепта;

– определение множеств $M_i(W_2)$, состоящих из претендентов на концепт, связываемый с концептом W_1 при помощи слова-связки L_i, удовлетворяющих установленному типу связи (для каждого L_i из множества $M(L_i)$);

– определение наиболее вероятного типа связи ($F_{(W_2,L_1)}\uparrow$ или $F_{(W_2,L_1)}\downarrow$) между будущим словом-связкой L_1 и концептом W_2 (для каждого L_i из множества $M(L_i)$);

– включение в онтологию наиболее вероятной связки из множества $M(T_i)$ возможных вариантов троек.

На заключительном этапе алгоритма определяется множество $M(T_i)$ – множество троек, для которых определены типы связей $F_{(W_i,L_1)}$ и $F_{(W_2,L_1)}$ ((2), (3), (4) или (5) соответственно). При этом предлагается ранжировать элементы из данного множества в соответствии со значениями вероятностей их выбора в качестве троек, актуальных для проектируемой онтологии.

Построение онтографа. Под онтографом, построенным по результатам выполнения предыдущих этапов, понимается двудольный граф, вершинами которого являются концепты высшего уровня и дочерние концепты онтологии предметной области, а дугами – связи между ними. Двудольный граф – это однонаправленный ориентированный граф, в одну вершину которого может входить и выходить несколько дуг.

Для машинного представления данных онтологии и последующего построения онтографа может быть использована реляционная база данных, включающая в себя связанные между собой таблицы концептов и связей между ними (для главного и дочерних уровней). При формировании онтографов на основе таблиц БД могут быть использованы редакторы онтологий (например, редактор Protege [7]). Предложенная модель создания онтологии поддерживает динамическое формирование структуры входящих в состав онтологии предметной области троек «концепт-связь-концепт», позволяя тем самым эффективно реализовать операции поэтапного формирования онтологии и редактирования ее структуры.

3. Программная реализация и оценка эффективности предложенного метода

По предложенному методу был разработан программный модуль «Concept-Ont-M2», который может эффективно использоваться для задач анализа электронных текстов и автоматического создания онтологий. Разработанный программный комплекс обеспечивает: автоматическое выделение наиболее значимых слов из корпуса текстов заданной предметной области с последующим их ранжированием по критериям значимости; формирование концептов-словосочетаний для высшего уровня онтологии; формирование троек «концепт-связка-концепт» с последующим выделением дочерних концептов-словосочетаний по методу главного концепта; построение онтографа, отражающего древовидную структуру сформированной онтологической модели. Хранение и коррекция динамически перестраиваемой концептуальной онтологической модели реализуется путем изменения состава и содержания записей, размещаемых в таблицах реляционной базы данных. Программный комплекс обеспечивает как ручной, так и автоматический ввод, редактирование и удаление слов и словосочетаний.

Ниже приведен пример построения по предложенному методу фрагмента онтологии предметной области «Информационные технологии». В качестве входного корпуса текстов использованы электронные тексты авторефератов диссертационных работ по специальности 05.13.06 – информационные технологии. В соответствии с описанным выше подходом выделяем и ранжируем множество концептов-словосочетаний, имеющих наибольшие значения $K(W_i)$. Используя формулы (6) – (10), выделяем тройки «концепт – связь – концепт».

Фрагмент ранжированного списка троек «главный концепт – связь – дочерний концепт» имеет вид:

«модели и информационные технологии администрирования информационного комплекса автоматизированных систем – название – диссертация»;

«модель актуализации оперативных данных – разработана в – диссертация»;

«объектная модель типового документа – усовершенствована в – диссертация»;

«математическая модель – которая описывает – информационный комплекс автоматизированной системы»;

«математическая модель – позволяет объединить – описание данных»;

«математическая модель – позволяет объединить – электронные документы»;

«электронные документы – объединяются по принципу – использование в конкретной функциональной задаче»;

«описание данных – объединяются по принципу – использование в конкретной функциональной задаче»;

«использование в конкретной функциональной задаче – позволяет расширить – функции администрирования»;

«функции администрирования – a part of – информационный комплекс автоматизированной системы»;

«функции администрирования – расширяются за счет – типизации функций»;

«математическая модель – получила дальнейшее развитие в – диссертация»;

«модель запроса – получила дальнейшее развитие в – диссертация»;

«модель запроса – относящийся к – информационный комплекс автоматизированной системы»;

«входные документы – is a – электронные документы»;

«входные документы – воздействуют на состояние – информационный комплекс»;
«модель актуализации оперативных данных – is a – модель актуализации оперативных данных информационного комплекса»;

«информационный комплекс – is a – информационный комплекс автоматизированной системы»;

«модель актуализации оперативных данных – позволяет автоматизировать – функция обработки документов»;

«модель актуализации оперативных данных – позволяет получить – количественная оценка фактографических данных»;

«объектная модель типового документа – представляет – электронные документы»;
«электронные документы – представляются как – иерархия модифицированных фреймов».

После ранжирования списка троек «концепт – связь – концепт» для каждого главного концепта строится корневое дерево связанных с ним концептов, в котором корнем является главный концепт, вершинами на нижних уровнях – связанные с ним концепты, а дугами – связи между концептами. Связи между концептами могут быть традиционными (is-a, a-part-of), а могут представлять собой глагольные группы, определенные при помощи метода главного концепта.

На рис. 1 представлен фрагмент построения корневых деревьев, корнями которых являются главные концепты, вершинами – связанные концепты, а дугами – связи между ними (наименования концептов и связей приведены в сокращенном виде – полностью открываются с помощью кликов). Если некоторые концепты присутствуют в нескольких деревьях, то при объединении деревьев повторяющиеся вершины отождествляются.

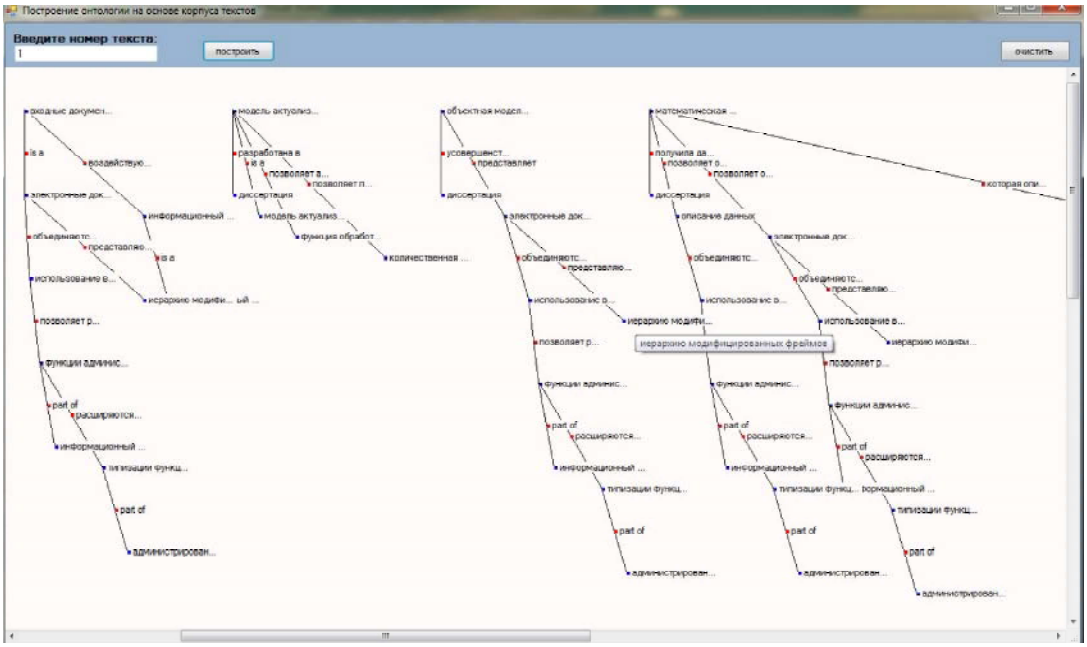

Рис.1

Для построения онтологии документа все построенные деревья объединяются в один онтограф (который может и не быть деревом). На рис.2 представлен фрагмент онтографа, построенного с учетом связей по методу главного концепта для одного текста. В построенном онтографе отождествлены идентичные вершины, соответствующие совпадающим концептам.

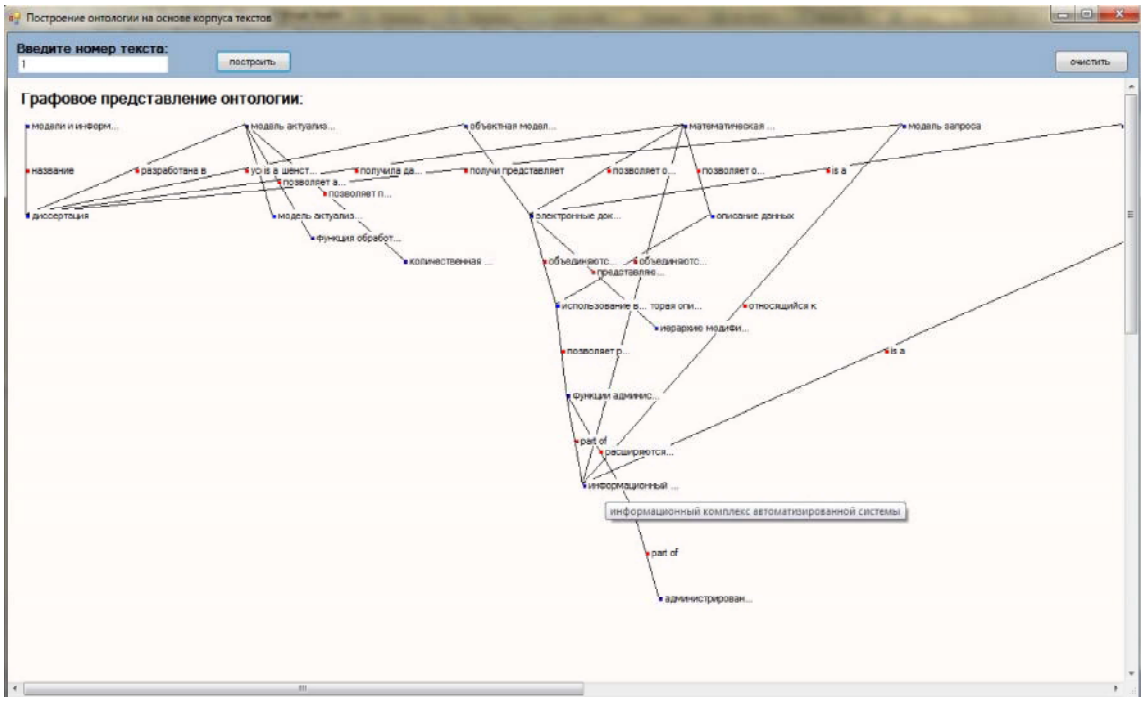

Рис. 2

Проведенные экспериментальные исследования показали, что онтологические модели, построенные с применением предложенного подхода (в частности, метода главного концепта) являются более представительными, чем модели, основанные на поиске отдельных связей между концептами. Оценка эффективности проводилась по двум параметрам: R – точность поиска связей (отношение правильно найденных связей к общему количеству найденных связей); P – полнота поиска связей (отношение правильно найденных связей к общему количеству связей, выявленных экспертом). Результаты экспериментальных исследований (для корпуса текстов из электронной библиотеки авторефератов по технической тематике): для метода поиска отдельных связей и метода главного концепта значения R составляют 59 и 83% соответственно; значения P – 79 и 86% соответственно.

4. Выводы и перспективы дальнейших исследований

Проведенные исследования позволяют сделать вывод, что важным этапом автоматического построения онтологий является выявление релевантных связей между концептами-словосочетаниями с последующим формированием троек «концепт – связь – концепт». Модификация и программная реализация метода нахождения таких связей с учетом расположения элементов «концепт-связка» в тексте позволили повысить эффективность автоматического построения онтологических моделей в виде онтографа. Вспомогательным этапом такого построения есть формирование корневого дерева, в котором корнем является главный концепт, вершинами на нижних уровнях – связанные с ним концепты, а дугами – связи между концептами. Если некоторые концепты присутствуют в нескольких деревьях, то при объединении деревьев повторяющиеся вершины отождествляются.

Научная новизна предложенного метода состоит в реализации оригинальной процедуры автоматического построения древовидной структуры концептов формируемой онтологии в рамках анализируемого корпуса текстов. При проведении дальнейших исследований целесообразно усовершенствовать предложенный метод, дополнив его анализом более сложных типов связок в онтологической модели и учетом дополнительных атрибутов формируемых онтологий.

Список л**итературы: 1.** *Палагин А.В.* Онтологические методы и средства обработки предметных знаний: Монография / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. Луганск: изд-во ВНУ им. В. Даля, 2012. 324 с. **2.** *Норенков И.П.* Интеллектуальные технологии на базе онтологий / И.П. Норенков // Информационные технологии. 2010. №1. С.17-23. **3.** *Антонов И.В.* Методы анализа данных в задачах автоматизации построения онтологии предметной области / И.В. Антонов, М.В. Воронов // Дистанционное и виртуальное обучение. 2011. N 8. С. 19-35. **4.** *Палагін О.В.* Розбудова абстрактної моделі мовно-онтологічної інформаційної системи / О.В. Палагін, М.Г. Петренко // Математичні машини і системи. 2007. №1. С. 42–50. **5.** *Чалая Л. Э.* Меры важности концептов в семантической сети онтологической базы знаний [Текст] / Л. Э. Чалая, Ю. Ю. Шевякова, А. Ю. Шафроненко// Матеріали другої міжнар. наук.-техн. конф. «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління». Киев : КДАВТ, 2011. С. 51. **6.** *Чалая Л. Э.* Определение значимости связей между концептами при автоматическом синтезе онтологий / Л. Э. Чалая, А. В. Чижевский // International Scientific Journal «Acta Universitatis Pontica Euxinus» (Special number). Varna. 2014. P. 480–484.

Поступила в редколлегию 17.12.2015

Чалая Лариса Эрнестовна, канд. техн. наук, доцент кафедры искусственного интеллекта ХНУРЭ. Адрес: Украина, 61166, Харьков, пр. Науки, 14, тел. 057-7021337, e-mail: larysa.chala@nure.ua.

Чижевский Антон Валериевич, аспирант кафедры искусственного интеллекта ХНУРЭ. Адрес: Украина, 61166, Харьков, пр. Науки, 14, тел. 057-7021337, e-mail: chij7@mail.ru