СОПОСТАВЛЕНИЕ СЕМАНТИЧЕСКИХ ИНФОРМАЦИОННЫХ РЕСУРСОВ WEB НА ОСНОВЕ ОНТОЛОГИЧЕСКОГО АНАЛИЗА

Юлия Рогушина

Аннотация*: В работе проанализированы современные средства представления распределенных знаний. Рассмотрены проблемы, возникающие в процессе управления такими знаниями в интеллектуальных информационных системах, ориентированных на работу в Web. Обоснована необходимость использования онтологий и актуальность развития семантических технологий, направленных на их поддержку, в частности, стандарты и языки Semantic Web. Рассматривается задача сопоставления онтологий, результаты такого сопоставления и частный случай этой задачи, когда сравниваются онтологии, являющиеся развитием одной и той же начальной онтологии. Рассмотрены семантические Wiki-ресурсы, как источник информации для пополнения и усовершенствования онтологий предметных областей, отображаемых с помощью набора Wiki-страниц. Проведен анализ компонентов онтологий предметных областей и семантической разметки Wiki-ресурсов (на примере Semantic MediaWiki), предложен метод их автоматизированного сопоставления для усовершенствования баз знаний распределенных интеллектуальных систем.*

Ключевые слова*: онтология предметной области, автоматическая генерация онтологий, онтология, Wiki-технологии.*

ITHEA Keywords*: I.2.4 Knowledge Representation Formalisms and Methods.*

Введение

Использование распределенных знаний является одним из определяющих факторов эффективности современных информационных систем. Это в свою очередь обуславливает актуальность исследований, направленных на приобретение и накопление знаний, решение проблем распознавание, логического вывода новых знаний для поддержки принятия решений

Интеллектуальные информационные системы (ИИС), которые работают в открытом распределенном информационном пространстве, требуют постоянного пополнения и

обновления знаний, которые поступают из внешней среды, так как в ИИС использование знаний о предметной области (ПрО) и методы их обработки играют решающую роль и определяют полезность получаемых результатов.

Из-за сложности получения знаний важное значение приобретает проблема обеспечения их интероперабельности и повторного использования.

В настоящее время для интероперабельного представления знаний в ИИС широко применяются онтологии, которые обеспечивают повторное использование полученных знаний в различных приложениях [Gruber, 1991]. Онтологии базируются на фундаментальном теоретическом базисе дескриптивных логик, для них уже существуют общепринятые стандарты описания, языки и программные средства. Особый интерес представляет применение онтологического анализа как основы для обработки распределенных знаний. В связи с этим значительное количество исследований связано с теоретическим базисом онтологий, их построением, усовершенствованием, с получением знаний из онтологических структур, а также с другими важными аспектами менеджмента онтологий, которые значительно различаются по цели и назначению такого анализа [Uschold, 1996]. Одним из важных направлений таких исследований является интеграция онтологий с другими информационными ресурсами Web. Такие свойства Web, как гетерогенность и динамичность, обуславливают ряд проблем, связанных с пополнением, использованием и оценкой онтологий.

Управление знаниями в Web

Проблема управления знаниями, которые являются составляющей различных ИИС, связана с интеллектуальным анализом данных и распознаванием свойств информационных объектов (ИО), которые используются и обрабатываются в предметной области, интересующей пользователя. В связи с тем, что современные ИИС используют и генерируют большие по объему и сложные по структуре совокупности знаний, возникает проблема разработки эффективных методов управления ими [Holtshouse, 2013].

Понятие «управление знаниями» («knowledge management») было введено американским ученым Карлом Виигом по аналогии к таким терминами, как «управление данными» и «управление информацией». Цель управления знаниями заключается в том, чтобы преобразовать процесс создания баз знаний из искусства в инженерную дисциплину. Это требует разработки соответствующих методов, языков и инструментов [Studer, 1998], в частности, связанных с построением и использованием онтологий. *Управление знаниями*

представляет собой совокупность процессов, связанных с эффективным созданием, сохранением, распространением и использованием знаний для достижения поставленных целей.

В процессе управления знаниями в Web необходимо решать такие проблемы, как:

- интеграция гетерогенных знаний, полученных из разных источников (например, интеграция онтологий или сопоставление различных видов семантической разметки, поиск по метаописаниям);
- поиск противоречий между знаниями, оценка их достоверности и надежности (например, сравнение знаний, извлеченных из контента различных ИР, сравнение двух онтологий);
- получение новых знаний из уже имеющихся (например, с помощью дедуктивного, традуктивного или индуктивного логического вывода) и их представление в форме, понятной пользователю (например, визуализация онтологических структур или семантических сетей);
- поиск знаний, необходимых конкретному пользователю (например, персонифицированный поиск, учитывающий знания о пользователе и его задаче, или семантический поиск сложных информационных объектов и сервисов);
- автоматизация создания матаданных, которые корректно отображают контент ИР (например, семантическая разметка ИР, – как текстовых, так и мультимедийных) на уровне содержания, и эффективный поиск в таких метаописаниях.

Любая ИИС может поддерживать определенное подмножество функций управления знаниями. Для Web-ориентированных ИИС эта задача усложняется тем, что нужно осуществлять управление распределенными и гетерогенными знаниями [Гладун, 2016]. Следует учитывать динамичность и гетерогенность среды Web: динамичность требует постоянного обновления знаний о ПрО, которые необходимо извлекать из открытых информационных ресурсов (ИР), а гетерогенность – интеграции полученной информации и ее проверки на непротиворечивость. Таким образом, возникает необходимость в средствах пополнения и усовершенствования онтологий путем анализа ИР. Не менее важная задача – обнаружить различия в онтологиях, описывающих одну и ту же ПрО. Это позволяет формализованно описать несоответствия в подходах различных пользователей, и, если надо, разрешить их.

В отличие от традиционных ИИС, те системы, которые ориентированы на работу в Web, для своего функционирования нуждаются в получения информации из внешних ИР. Доступ к некоторым из них ИИС может получать непосредственно, но в большинстве случаев в качестве

посредников используются внешние информационно-поисковые системы (ИПС). Однако внешние ИПС позволяют лишь построить множество ИР, из которых можно получить информацию, и не решают проблему извлечения нужных знаний из этих ресурсов [Гладун, 2013].

Одним из наиболее известных направлений в исследованиях управления знаниями в Web является проект Semantic Web, предложенный Тимом Бернес-Ли. Основная идея этого проекта заключается в преобразовании Web в глобальную базу знаний и разработке соответствующих стандартов, языков и программных средств для обработки информации в этих ресурсах на семантическом уровне [Berners-Lee, 2006] .

Анализ современных исследований в области управления знаниями показывает, что, несмотря на значительные достижения в сфере Data Mining и Text Mining, извлечение знаний из естественно-языковых неструктурированных ИР остается крайне трудоемкой и полностью не автоматизируемой задачей. Значительно более эффективно осуществляется извлечение знаний из ИР, содержащих семантическую разметку, которую можно сопоставить с теми или иными структурами формализованного представления знаний.

Для достижения практически полезных результатов необходимо ориентироваться на те средства семантической разметки, которые уже сегодня получили достаточно широкое распространение, имеют достаточно высокую выразительную мощность, надежную реализацию и удобный пользовательский интерфейс. Этим требованиям удовлетворяют семантические Wiki, в частности, Semantic MediaWiki [Völkel, 2006]: эта платформа сегодня широко используется, сформированные на ее основе ИР часто обновляются и быстро растут. Использование таких источников обуславливается тем, что Wiki-ресурсы динамически обновляются всем сообществом пользователей, имеют четко определённую и простую для понимания структуру и обеспечивают обработку информации на семантическом уровне, и, таким образом, обеспечивают технологическую платформу для группового управления знаниями [Majchrzak, 2006] .

Встроенные средства Semantic MediaWiki позволяют строить онтологии, представленные на языке OWL, но эти средства недостаточно гибки для того, чтобы построенную с их помощью онтологию можно было интероперабельно применять в других ИИС. Поэтому возникает необходимость в разработке более совершенных методов формирования и пополнения онтологий на основе Wiki-ресурсов, а также сравнения онтологий, сформированных и усовершенствованных на основе различных состояний Wiki-ресурсов.

Постановка задачи

Чтобы обеспечить интеллектуальные информационные системы возможностью импортировать, модифицировать и обновлять онтологические знания, извлекаемые из внешних семантически размеченных Wiki-ресурсов, необходимо разработать методы и средства, обеспечивающие решение следующих задач:

- построить взаимно-однозначное соответствия между основными элементами онтологии ПрО (классы и подклассы, экземпляры, свойства объектов и данных, их значение) и Semantic MediaWiki (категории, семантические свойства, их значения и связи);
- разработать средства и критерии сопоставления онтологий, являющихся усовершенствованиями одной и той же онтологии ПрО, пополненной знаниями из различных Wiki-ресурсов, семантически размеченных элементами начальной онтологии.

На первом этапе необходимо построить (или выбрать) начальную онтологическую модель ПрО, которая будет использоваться для семантической разметки. Для этого следует разработать метод преобразования элементов онтологии в конструкции Semantic MediaWiki (категории и семантические свойства). На следующем этапе необходимо обеспечить возможность уточнения и усовершенствования начальной онтологии ПрО на основе анализа семантически размеченных Wiki-ресурсов, то есть разработать метод преобразования конструкций Semantic MediaWiki в онтологию, представленную на языке OWL. Затем следует проанализировать полученную онтологию, оценить ее свойства и соответствие представлениям пользователя о ПрО. Итеративное повторение этих действий должно обеспечить формирование адекватной онтологии ПрО, которую можно использовать в различных ИИС в качестве базы знаний.

Основные семантические компоненты Semantic MediaWiki

Semantic MediaWiki (SMW) позволяет пользователям добавлять семантические аннотации к Wiki-страницам, используя дополнительные элементы разметки, превращая MediaWiki в семантический ресурс: категории, семантические свойства и запросы. SMW использует концепцию семантических свойств (для создания данных) и семантических запросов (для использования данных). Пользователи размечают статьи категориями и свойствами, чтобы информация становилась доступной для запросов. Формальная модель Wiki-ресурса состоит из следующих элементов [Rogushina, 2016]:

- множества Wiki-страниц $P = P_{user} \cup P_{categ} \cup P_{template} \cup P_{spec}$, где P_{user} – множество страниц, созданных пользователями, P_{categ} – множество страниц, которые описывают категории, $P_{template}$ – множество страниц, которые описывают шаблоны, P_{spec} – множество других специальных страниц;
- $L = \{"link"\}$ множество из одного элемента, который описывает ссылку одной Wiki-страницы этого ресурса на другую (хотя в Wiki-ресурсах предусмотрены и ссылки на другие виды страниц, в рамках данной модели они не учитываются).

Семантические свойства обеспечивают привязывание данных к Wiki-страницам. Каждое свойство имеет тип, название и значение, кроме того, ему соответствует отдельная Wiki-страница в специальном пространстве имен, которая позволяет задавать тип свойства, определять его положение в иерархии свойств, а также документировать его использование. Свойства, в отличие от категорий MediaWiki, имеют не только название, но и значение. В текст Wiki-страницы семантическое свойство вставляется в формате [[Имя свойства::Значение свойства]]. Семантические свойства используются для того, чтобы определить смысловую нагрузку ссылок на другие страницы.

Формальная модель семантически обогащенных Wiki-ресурсов является более сложной и включает ряд элементов, связанных с семантическими свойствами [Rogushina, 2016] :

$$W_s = \left\langle P = P_{user} \cup P_{categ} \cup P_{template} \cup P_{sem\_prop} \cup P_{spec}, L = \{"link"\} \cup L_{sem\_prop} = \{l_i\} \right\rangle \qquad (1)$$

где P_{sem_prop} – множество страниц, которые описывают семантические свойства Wiki-страниц, некоторые из которых являются семантически определенными ссылками на другие Wiki-страницы: $P_{sem\_prop\_page} \subseteq P_{sem\_prop}$, а другие связывают страницы со значениями различных типов данных (эти типы данных определяются на страницах семантических свойств).

Основные семантические компоненты онтологий

Онтологический анализ позволяет превращать описание определенного представления о внешнем мире в набор терминов и правил их использования, пригодных для машинной обработки. Онтология представляет собой явную спецификацию концептуализации. Онтологический подход поволяет анализировать окружающий мир независимо от формы представления знаний о нем [Guarino, 1995]. Онтологию можно рассматривать как базу знаний (БЗ) специального вида с семантической информацией об определенной ПрО. Компоненты, из которых складываются конкретные онтологии, зависят от парадигмы представления, но практически все модели онтологий содержат определенные концепты (понятие, классы), свойства концептов (атрибуты, роли), отношение между концептами (зависимости, функции) и ограничения использования, которые определяются аксиомами. В общем случае формальная модель онтологии O – это упорядоченная тройка O=<T,R,F>, где T – множество понятий ПрО; R – множество отношений между ними; F – множество функций интерпретации понятий и отношений. Эта формальная модель может быть конкретизирована в зависимости от назначения и сферы применения онтологии.

Сейчас в ИИС для представления онтологий чаще всего используются различные диалекты языка OWL. Онтология OWL (Web Ontology Language) – это последовательностью аксиом и фактов, а также ссылок на другие онтологии.

Фундаментальные понятия определенной ПрО должны соответствовать *классам* онтологии. Каждый экземпляр в онтологии OWL принадлежит к классу owl:Thing, а каждый созданный пользователем класс автоматически является подклассом owl:Thing. Специфичные для ПрО корневые классы определяются простым объявлением именованного класса. OWL также позволяет задать пустой класс: owl:Nothing. Определения могут быть расширяемыми и распределенными. Фундаментальным таксономическим конструктором для классов является rdfs:subClassOf. Он связывает отдельный класс с общим. Если X – подкласс Y, то каждый представитель X – также представитель Y. Отношение rdfs:subClassOf является транзитивным. Если X – подкласс Y и Y – подкласс Z, тогда X – подкласс Z.

Определение класса состоит из двух частей: названия (или ссылки на него) и списка ограничений. Каждое выражение, которое непосредственно помещается в определении класса, уточняет свойства представителей этого класса. Представители класса принадлежат пересечению указанных ограничений. Для определения *экземпляра* достаточно объявить его членом какого-либо класса.

Свойства дают возможность утверждать общие факты о членах классов и об экземплярах. Они представляют собой бинарные отношения. Различают отношения между представителями классов и RDF-литералами или типами данных, определенными XML Schema, и отношения между представителями классов.

При определении свойства существует много способов ограничить это отношение. Можно задать его домен и диапазон. Свойство может быть определено как специализация (подствойство) уже существующего свойства. Возможны и более сложные ограничения. Свойства, также как и классы, могут быть организованы иерархически. Формальная семантика OWL содержит описание того, как получить логические следствия, имея такую онтологию, то есть получить факты, которые не представлены в ней непосредственно, но тем не менее обусловлены ее семантикой.

Семантику онтологических языков обычно раскрывают через теорию моделей. В частности, она определяет функцию интерпретации, которая проецирует каждый элемент онтологии на множество D, которое называют областью интерпретации.

Формальная онтология поволяет анализировать окружающий мир независимо от формы представления знаний о нем. Формальные модели онтологий позволяют отображать различные метасвойства их элементов онтологий, такие как идентичность (identity), стойкость (rigidity), согласованность (unity) и зависимость (dependence), обеспечивая возможность онтологического концептуального анализа ПрО [Guarino, 2000]. Анализ различных средств представления онтологий и их формальных моделей, которые предлагают технологии Semantic Web, показывает, что они значительно отличаются своими выразительными возможностями и своей сложностью: RDF Schemas предлагает простейший уровень для представления онтологий, а OWL Full – наиболее сложный. Выбор средства представление онтологии зависит от специфики проблемы, для которой она разрабатывается. Например, для семантического поиска обычно достаточно простейшего уровня представления знаний, тем не менее некоторые задачи, связанные с нахождением совокупности ИО со сложной структурой и многочисленными условиями, требуют более сложных возможностей для отображения знаний соответствующих ПрО.

Будем использовать формальную модель

$$O = \left\langle X = X_{cl} \cup X_{ind}, R = r_0 \cup \{r_i\} \cup \{p_j\}, F, T \right\rangle \qquad (2)$$

Эта модель состоит из таких элементов:

- $X = X_{cl} \cup X_{ind}$ – множество концептов онтологии, где X_{cl} – множество классов, X_{ind} – множество экземпляров классов, таких, что $\forall a \in X_{ind} \exists A \in X_{cl}, a \in A$;
- $R = r_{ier\_cl} \cup \{r_i\} \cup \{p_j\}$ – множество отношений между элементами онтологии, где r_{ier_cl} – иерархическое отношение, которое может устанавливаться между классами онтологии и свойствами классов и характеризуется такими свойствами, как антисимметричность и транзитивность, $r_{ier\_cl} : X_{cl} \to X_{cl}$; $\{r_i\}$ – множество объектных свойств, которые устанавливают отношение между экземплярами классов: $r_i(a, a \in X_{ind}) = b, b \in X_{ind}$, $r_i : X_{ind} \to X_{ind}$; $\{p_i\}$ – множество свойств данных, которые устанавливают отношения между экземплярами классов и значениями: $p_i(a, a \in X_{ind}) = t, t \in T$, $p_i : X_{ind} \to Const$, причем внутри множеств объектных свойств и свойств отношений также могут существовать иерархические отношения r_{ier_obj}, $r_{ier\_obj} : \{r_i\} \to \{r_i\}$ и r_{ier_data}, $r_{ier\_data} : \{p_j\} \to \{p_j\}$;
- F – множество свойств классов онтологии, экземпляров классов и их свойств, которые могут применяться для логического вывода (например, эквивалентность, отличие);
- T – множество типов данных (например, строка, целое).

Выбор именно такой модели онтологии обуславливается следующими причинами. Во-первых, она наиболее полно соответствует интуитивному представлению об онтологиях, заложенному в пользовательском интерфейсе редактора онтологий Protégé [Tudorache, 2013]. Во-вторых, эта модель достаточно легко интегрируется с моделью тезауруса задачи, который используется для семантического поиска. В-третьих, эту модель позволяет сопоставить онтологическое представление знаний о ПрО с семантическими конструкциями Semantic MediaWiki: для некоторых элементов может быть сформировано взаимно-однозначное соответствие, а сопоставление других требует дополнительных преобразований, но также может быть хотя бы частично автоматизировано.

Рассмотрим подробнее эти соответствия. Для этого надо проанализировать основные семантические элементы Semantic MediaWiki. Следует отметить, что, хотя в Semantic MediaWiki предусмотрено автоматическое построение онтологий по семантически размеченным Wiki-страницам, но в текущих версиях это построение учитывает слишком мало семантических параметров, а его результаты мало пригодны для дальнейшего использования.

Важным аспектом предложенной модели является возможность ее интеграции с формальной моделью тезауруса задачи, который используется для семантического поиска.

Применение онтологического анализа в семантической разметке

При обработке информации на естественном языке возникает необходимость в использовании знаний о ПрО, для чего часто используются специализированные онтологии. Основное их назначение в таких задачах – обеспечить связь между фрагментами текста на естественном языке (ЕЯ) и понятиями предметной области (например, классами или экземплярами онтологии). В частности, широко используются тезаурусы и лингвистические онтологии [Loukachevitch, 2014].

Сегодня наиболее распространенными подходами к представлению лингвистических знаний являются традиционные информационно-поисковые тезаурусы, тезаурус WordNet [Pedersen, 2004] и специализированные формальные онтологии.

Тезаурус – это словарь основных понятий языка, которые выражаются отдельными словами или словосочетаниями, с определенными семантическими связями между ними. Как правило, тезаурусы используются для семантической обработки естественно-языковых ИР, например, при семантическом поиске или при семантической разметки. С тезаурусами тесно связаны лексические онтологии, которые представляют собой БЗ онтологического типа о понятийной системе и лексико-терминологическом составе ПрО.

Семантическая разметка связывает текст X с набором тегов разметки при помощи соответствующего языка разметки. Под текстом X, $X = \langle x_1,...,x_n \rangle$ будем понимать конечную непустую последовательность символов $x_i \in A, i = \overline{1,n}$, разделенную на отдельные слова разделителями, из конечного множества B, $B \subseteq A$.

Разметка связывает произвольный фрагмент текста $\langle x_p,...,x_{p+q} \rangle, 1 \le p \le n, 1 \le p+q \le n, q \ge 1$ с тэгом $t_i \in T$. Тэги разметки – элементы конечного множества T, $T = \{t_1,...,t_m\}$. Наборы тэгов может формироваться из разных источников в зависимости от того, какой язык разметки используется и какая ПрО должна быть охарактеризована. Примеры языков разметки: SGML, RDF Schema, тэги микроформатов, элементы Dublin Core , множество терминов произвольной онтологии.

Разметку называют *семантической*, если для элементов множества тэгов T описана их семантика. Выразительные возможности семантической разметки в значительной мере зависят от набора тэгов, которые используются для этой разметки, их объема, видов отношений между этими тэгами и возможностей логического вывода на этих структурах. Наибольшую

выразительность и универсальность как источник тэгов для семантической разметки имеют онтологии: если в качестве тэгов используются термины онтологии, то это дает возможность установить связь между фрагментами этого текста и понятиями ПрО, которая отображена в онтологии, то есть определить семантическую структуру этого текста [Vargas-Vera, 2002] .

Для автоматизированного выполнения и использования семантической разметки целесообразно использовать специализированную *лексическую онтологию* (ЛО) [Лесько, 2010]. Лексическая онтология – это легкая онтология, которая содержит два основных класса: «понятие ПрО» и «словоформа». Между экземплярами этих классов существует отношение «соответствует». Для каждого экземпляра класса «понятие ПрО» должен существовать хотя бы один экземпляр класса «словоформа», что находится с ним в отношении «соответствует»: $\forall t_i \in T \exists l_{i_1} \in$'словоформа "$= t_i$. Принципиальное отличие ЛО от обычных лексических онтологий заключается в том, что она строится для конкретной онтологии ПрО и поэтому как экземпляры класса «лексема» в нее включаются только те понятия, которые связаны с понятиями соответствующей онтологии, а не все понятия, которые присутствуют в естественном языке. Это значительно уменьшает объем ЛО и, соответственно, сокращает время обработки текста.

Задача сопоставления онтологий

Методы согласования онтологий помогают пользователям находить сходства и расхождения между онтологиями или интегрировать две онтологии в результирующую онтологию, которая содержит элементы обеих начальных онтологий. Для этого необходимо определить соответствия между концептами в онтологиях, обеспечить операции отображения, выравнивание и объединение. *Отображение* (сопоставление) онтологий (ontology mapping) – установление соответствия между несколькими онтологиями. *Выравнивание* онтологий (ontology alignment) – установление разных видов соответствия между онтологиями. *Интеграция* онтологий (ontology merging) – создание новой согласованной онтологии или фрагмента онтологии с двух или больше имеющихся онтологий.

При сопоставлении онтологий возникают проблемы их неоднородности. *Синтаксическая* неоднородность – онтологии построены на разных языках представления знаний или с использованием разных формализмов представление знаний, например, с помощью OWL и F-логики. *Терминологическая* неоднородность – расхождения в именах, которые ссылаются на одни и те же сущности в разных онтологиях. *Концептуальная* (семантическая) неоднородность – расхождение в моделировании одной и той же ПрО из-за использования

разных (хотя, возможно, эквивалентных) аксиом для определения концептов или же вследствие использования совсем разных понятий.

Различия между онтологиями могут быть обусловлены целью их создания и критерием классификации объектов и не являются ошибкой или противоречием, а характеризуют представления относительно ПрО разных пользователей и в разные моменты времени. Поэтому в многих случаях самое нахождения таких отличий является важным результатом .

Проблема сопоставления онтологий обусловлена тем, что:

- элементы онтологии (классы, свойства, связи, объекты) с одинаковыми именами могут иметь разную семантику;
- элементы онтологии (классы, свойства, связи, объекты) с одинаковой семантикой могут иметь разные имена.

Методы сопоставления онтологий подразделяют на лингвистические, структурные, статистические и семантические.

Лингвистический анализ определяет сходство между сущностями на основе сравнения их имен и анализа синонимии. Такой вид анализа является начальным этапом установления соответствий между сущностями онтологий.

Структурный анализ сравнивает сходства связей между сущностями (в частности и иерархическими). Оценка схожести двух сущностей двух онтологий может базироваться на позициях этих сущностей в иерархии классов: если подклассы и надкласс двух сущностей двух онтологий подобны, то и сами такие сущности тоже могут быть подобными. Анализ сходства по перекрестным связям для определения сходства между сущностями проводится так: если класс A1 связан с классом B1 связью типа R1 в одной онтологии, а класс A2 связан с B2 связью типа R2 в другой онтологии, и если известно, что B1 и B2 – похожи и R1 и R2 – похожи, то можно предположить схожесть A1 и A2. Так же можно говорить и о сходствах типов связей между R1 и R2, если известно, что A1 и A2 – похожи и B1 и B2 – похожи. Например, если известно, что классы «Человек» и «Лицо» – подобны и отношение «работать» и «быть сотрудником» тоже подобны, причем в первой онтологии класс «Человек» связан отношением «работать» с классом «Организация», а во второй – класс «Лицо» связан отношением «быть сотрудником» с классом «Компания» и «работать», тогда классы разных онтологий «Организация» и «Компания» являются подобными.

Статистический анализ базируется на использовании имеющихся экземпляров двух классов для оценки экстенсионального соответствия этих классов. Для нахождения соответствия между сущностями используются такие диагностические правила:

- класс C1 эквивалентен классу C2, если невозможно найти такой экземпляр O1 класса C1, который бы не принадлежал к классу C2, и наоборот;
- класс C1 является подклассом класса C2, если невозможно найти экземпляр O1 класса C1, который бы не принадлежал к классу C2, и класс C1 не эквивалентен классу C2.

Логический анализ выявляет надклассы для тех классов, которые сопоставляются, и анализирует наложенные на них ограничение. Например, в одной онтологии может быть класс «Ребенок», который является видовым классом для класса «Человек» с ограничением на свойство «Возраст»<16, а другая онтология содержит класс «Несовершеннолетнее лицо», который является видовым классом для класса «Лицо» с ограничением на «Возраст лица»<18. Вследствие анализа соответствия между классами «Ребенок» и «Несовершеннолетнее лицо» обнаруживаются родительные классы «Человек» и «Лицо». По наличию информации о соответствии этих классов осуществляется сравнение ограничений, наложенных на эти родительные классы. Для этого сравниваются свойства классов «Возраст» и «Возраст лица»: если эти свойства похожи, то проводится сравнение наложенных ограничений «<16» и «<18». Следствием такого сравнения является вывод, что класс «Ребенок» является подклассом класса «Несовершеннолетнее лицо». После получения локальных соответствий между сущностями можно определять глобальное соответствие между сущностями.

Инструментальные средства поиска соответствия между онтологиями классифицируют по назначению:

- объединение онтологий для создания новой (PROMPT, Chimaera, OntoMerge);
- определение функции преобразование из одной онтологии на другую (OntoMorph);
- обнаружение пар подобных концептов в онтологиях (OBSERVER, FCA-Merge);
- определение правил отображения для связи релевантных частей онтологий (ONION).

Следует отметить, что все эти средства позволяют только строить предположения о наличии соответствий, которые могут оказаться как истинными, так и ложными.

Задача сопоставления (сравнения) терминов онтологий является начальным этапом для построения гипотез относительно сходства между структурными элементами онтологий и требует дальнейшей обработки и проверки пользователем. Неверное сопоставление может быть результатом того, что в близких ПрО используется похожая терминология, но сами

термины имеют разное значение. В таких случаях экземпляры могут иметь свойства, а классы – подклассы и сверхклассы с подобными названиями, но с различной семантикой.

Первый этап сопоставления терминов онтологий можно выполнить несколькими способами:

- если имена этих элементов тождественны;
- если имена элементов семантически тождественны, например, являются синонимами или переводом на другой язык (такое сопоставление может выполняться автоматизированно с помощью лингвистических методов и находится вне рассмотрения в данной работе);
- непосредственно пользователем на основании его представлениям относительно специфики ПрО (например, синонимичные термины в рамках определенной ПрО);
- при помощи логического вывода из тождественности между другими элементами онтологий [Mishra, 2010] , т.е. путем анализа свойств классов и экземпляров классов этих онтологий, с помощью которых из имеющихся соответствий логически выводятся другие соответствия, либо путем анализа отношений между теми классами и экземплярами, для которых уже установленные соответствия.

Если результаты последнего способа всегда истинны (например, если установлено соответствие между классом $x_1 \in X_{cl_1}$ онтологии O_1 и классом $x_2 \in X_{cl_2}$ онтологии O_2 и если в онтологии O_1 содержится информация относительно того, что $x_1 \in X_{cl_1}$ эквивалентен классу $y_1 \in X_{cl_1}$, тогда можно считать, что установлено соответствие между классом $y_1 \in X_{cl_1}$ онтологии O_1 и классом $x_2 \in X_{cl_2}$ онтологии O_2) и такой вывод не требует дополнительного подтверждения от пользователя, то результаты, полученные остальными способами, имеют лишь вероятностную достоверность, и потому могут считаться истинными лишь после подтверждения пользователем. Например, если установлено соответствие между классами $x_1 \in X_{cl_1}$ онтологии O_1 и $x_2 \in X_{cl_2}$ онтологии O_2, и между классами $y_1 \in X_{cl_1}$ онтологии O_1 и $y_2 \in X_{cl2}$ онтологии O_2, а также в онтологии O_1 $x_1 \in X_{cl_1}$ является подклассом $y_1 \in X_{cl_1}$, тогда можно предположить, что $x_2 \in X_{cl_2}$ также является подклассом $y_2 \in X_{cl2}$. Тем не менее, онтология O_2 может связывать эти классы иным образом, в зависимости от критерия классификации.

Это подтверждает важность сведений о свойствах элементов онтологии (например, о тождественности классов), которые не имеют аналогов в семантических Wiki-ресурсах – таким образом, выразительная мощность онтологий значительно выше.

Если не удалось установить ни одного соответствия между элементами онтологий, по их сопоставление заканчивается полной неудачей, то есть считается, что представленные в них знание не пересекаются. В таком случае интеграция онтологий может выполняться путем их непосредственного объединения, которое заключается в объединении пространств имен всех элементов и отношений между ними.

Результатом сопоставления онтологий O_1 и O_2 может быть следующее:

- онтологии O_1 и O_2 тождественны;
- онтология O_1 является подмножеством онтологии O_2;
- онтологии O_1 и O_2 пересекаются, то есть имеют общую часть (на практике это обычно означает, что они происходят от одной онтологии);
- схемы онтологий O_1 и O_2 тождественны, но их экземпляры не совпадают (можно выделить случаи, в которых множества экземпляров пересекаются, не пересекаются и одно является подмножеством другого);
- схема онтологии O_1 является подмножеством схемы онтологии O_2;
- схемы онтологий O_1 и O_2 пересекаются;
- онтологии O_1 и O_2 отличаются, но количество отличий не превышает определенную количественную меру семантической близости между онтологиям, то есть можно назвать онтологии O_1 и O_2 версиями одной онтологии.

Для того, чтобы предлагать алгоритмы распознавания этих ситуаций, нужно предоставить их формальное определение.

Будем считать, что онтологии O_1 и O_2 *тождественны*, если существует однозначное взаимное сопоставление для всех классов, экземпляров классов, их свойств и их значений.

Будем считать, что онтологии O_1 и O_2 *имеют тождественные схемы*, если существует однозначное взаимное сопоставление для всех классов, их свойств и их значений.

С практической точки зрения наибольший интерес вызывает поиск расхождений между онтологиями и оценка семантической значимости этих расхождений. Другие важная проблема, связанная с найденными расхождениями, – это оценка возможности непротиворечащего объединения таких онтологий и нахождение их пересечения.

После того, как отличия между онтологиями найдены, возникает проблема количественной оценки этих отличий. Количественная оценка может выделять разные типы отличий – например, отличия в использовании иерархических отношений могут считаться более значительными по сравнению с отличиями в использовании отношений, специфичных для ПрО.

Кроме того, часто возникает необходимость сопоставления онтологий с другими семантическими структурами или с различными ИР, неявно содержащими соответствующие знания. Такие сопоставления в чем-то подобны правилам преобразования схем для интероперабельных баз данных. В частности, выделяют несколько видов сопоставления элементов онтологий [Studer, 1998]:

- сопоставление *переименования* (Renaming) используется для преобразования специфичных терминов ПрО в термины, связанные с методами;
- сопоставление *фильтрации* (Filtering) обеспечивает средства для выбора подмножества экземпляров ПрО для соответсвующего понятия метода;
- сопоставление класса (Class) предоставляет функции для вычисления экземпляров соответсвующего понятия метода из определений понятий в приложении, а не из экземпляров в приложении.

Еще одна важная задача, которая находится вне рамок классического управления знаниями, – это сопоставление онтологий с естественными тестами. Ее подзадачами являются разметка ЕЯ-текста терминами онтологии, пополнение онтологии знаниями, добытыми из размеченного ЕЯ-тексту, и вычисления меры семантической близости между текстом и онтологией. Решение этой задачи нуждается в учет специфики отдельных ЕЯ, и потому существующие средства и методы решения этой задачи должны разрабатываться для каждого языка в отдельности.

Построение лингвистической БЗ, которое позволяет соотносить фрагменты ЕЯ-текста с терминами онтологии, также находится вне рассмотрения данной работы (следует отметить, что на сегодня же существует определенное количество таких БЗ, в том числе и для украинского языка, и средств использования этих БЗ в семантической разметке, которые пригодны для использования в данной задаче). Кроме лингвистической БЗ, для сопоставления терминов онтологии с ЕЯ-текстом целесообразно использовать средства лингвистического анализа (например, для решения омонимии и определения частей предложение).

С точки зрения управление знаниями больший интерес вызывает обратная задача – по ЕЯ-тексту, который семантически размечен терминами онтологии, усовершенствовать эту онтологию знаниями, которые помещаются в этом тексте. Такое усовершенствование может обеспечить как установление новых связей между существующими элементами онтологии

(например, если фрагменты одного предложения связанные с экземпляром класса и , так и пополнение онтологии новыми элементами.

Сопоставление онтологий и семантических Wiki-ресурсов

Проблема сопоставления онтологий и семантических Wiki-ресурсов возникает в нескольких случаях. Во-первых, при создании семантических Wiki-ресурсов необходимо вначале сформировать набор категорий и семантических свойств. Но встроенные средства Semantic MediaWiki не позволяют ни визуализировать эту информацию, ни оценить ее целостность и непротиворечивость. Поэтому целесообразно вначале построить онтологию той ПрО, которая отображается в Semantic MediaWiki, а затем использовать эту онтологию в качестве основы для семантической разметки. Во-вторых, семантически размеченные Wiki-ресурсы намного более динамичны по сравнению с онтологиями – в их усовершенствовании и обновлении может участвовать широкий круг пользователей, и поэтому они могут быть полезны для усовершенствования соответствующей онтологии ПрО.

Рассмотрим более детально соответствия между элементами онтологии ПрО и страниц Semantic MediaWiki. Некоторые такие соответствия взаимно-однозначны и могут выявляться автоматически, некоторые – требуют дополнительных уточнений от пользователя. Для более строгого описания этих соответствий воспользуемся формальными моделями (1) и (2).

Table 1. Соответствия между основными элементами онтологий и Wiki-ресурсов

Semantic MediaWiki	Онтология	Отображение из Wiki в онтологию	Отображение из онтологии в Wiki
Категория	Класс	Взаимно-однозначное $P_{categ} \rightarrow X_{cl}$	Многозначное $X_{cl} \rightarrow P_{categ} \cup P_{template}$
Иерархия категорий	Иерархия классов	Многозначное	Взаимно-однозначное
Wiki-страница	Экземпляр класса	Многозначное $P_{user} \rightarrow X_{ind}$	Взаимно-однозначное $X_{ind} \rightarrow P_{user}$
Ссылка на другую Wiki-страницу	Объектное отношение	Взаимно-однозначное $L = \{"link"\} \rightarrow R$	Взаимно-однозначное $R \rightarrow L = \{"link"\}$

Семантическое свойство типа «страница»	Объектное свойство	Взаимно-однозначное $P_{sem\_prop\_page} \rightarrow \{r_i\}$	Взаимно-однозначное $\{r_i\} \rightarrow P_{sem\_prop\_page}$
Семантическое свойство любого другого типа	Свойство данных	Взаимно-однозначное $P_{sem\_prop} \rightarrow \{p_i\}$	Взаимно-однозначное $\{p_i\} \rightarrow P_{sem\_prop}$
Шаблон	Класс	Взаимно-однозначное $P_{template} \rightarrow X_{cl}$	Многозначное $X_{cl} \rightarrow P_{categ} \cup P_{template}$

Представление значений семантических свойств отображается в лексическую онтологию: для пополнения ЛО словоформами используется информация из конструкции Semantic MediaWiki: [[семантическое свойство:: термин |словоформа]].

Таким образом, если уже построена онтология на языке OWL, то ее достаточно просто использовать в Semantic MediaWiki. Но обратный процесс не может быть полностью автоматизирован. Более того, при автоматической генерации онтологии по Semantic MediaWiki будут утеряны содержащиеся в OWL-онтологии сведения о характеристиках классов и свойств, которые не имеют аналогов Wiki (в частности, об эквивалентности классов и свойств, их непересекаемости, об их области значения и определения). В то же время, часть контента Semantic MediaWiki не может быть непосредственно трансформирована в онтологии. Например, тот факт, что в страницах использован одни и тот же шаблон, говорит о том, что на эти страницах описаны информационные объекты одного типа, но для отображения этого в онтологии надо создать специфичный класс и связать его с элементом страницы. Но затем по такой онтологии невозможно понять, что надо создать в Wiki – шаблон или категорию: выбор зависит от пользователя, т.к. для ИО со специфичной, но формализованной структурой целесообразно создавать шаблоны, а во всех остальных случаях – страницы, которые будут отнесены к какой-либо категории. Кроме того, нельзя связать с классом онтологии не всю страницу, а ее конкретный фрагмент (кроме тех случаев, когда у него есть подзаголовок).

Практическое использование

Предложенный подход к управлению знаниями был апробирован при разработке интеллектуальной системы информационного и когнитивного сопровождения функционирования Национальной рамки квалификаций, в которой онтологический анализ использовался для описания компетенций [Rogushina, 2012].

На основе анализа естественноязыковых описаний национальных и европейских рамок [Lundqvist, 2011] была построена онтология, которая отображала эталонную модель рамки квалификаций. Эта модель описывает семантические свойства и отношения ИО, которые связаны с результатами обучения. Онтология формализует отношение между этими ИО и устанавливает их иерархию. Пользователь семантически размечает контент Wiki-страниц элементами этой онтологии: определяет категорию (или набор категорий) любого документу и идентифицирует отдельные элементы его контенту (рис.1). Использование онтологической модели рамки квалификаций обеспечивает автоматизировать обработку результатов обучения и интегрируя знания, необходимые пользователям [Rogushina, 2016]. В частности, наличие онтологии позволяет пользователям представить структуру категорий и семантических свойств, которые можно использовать в запросах.

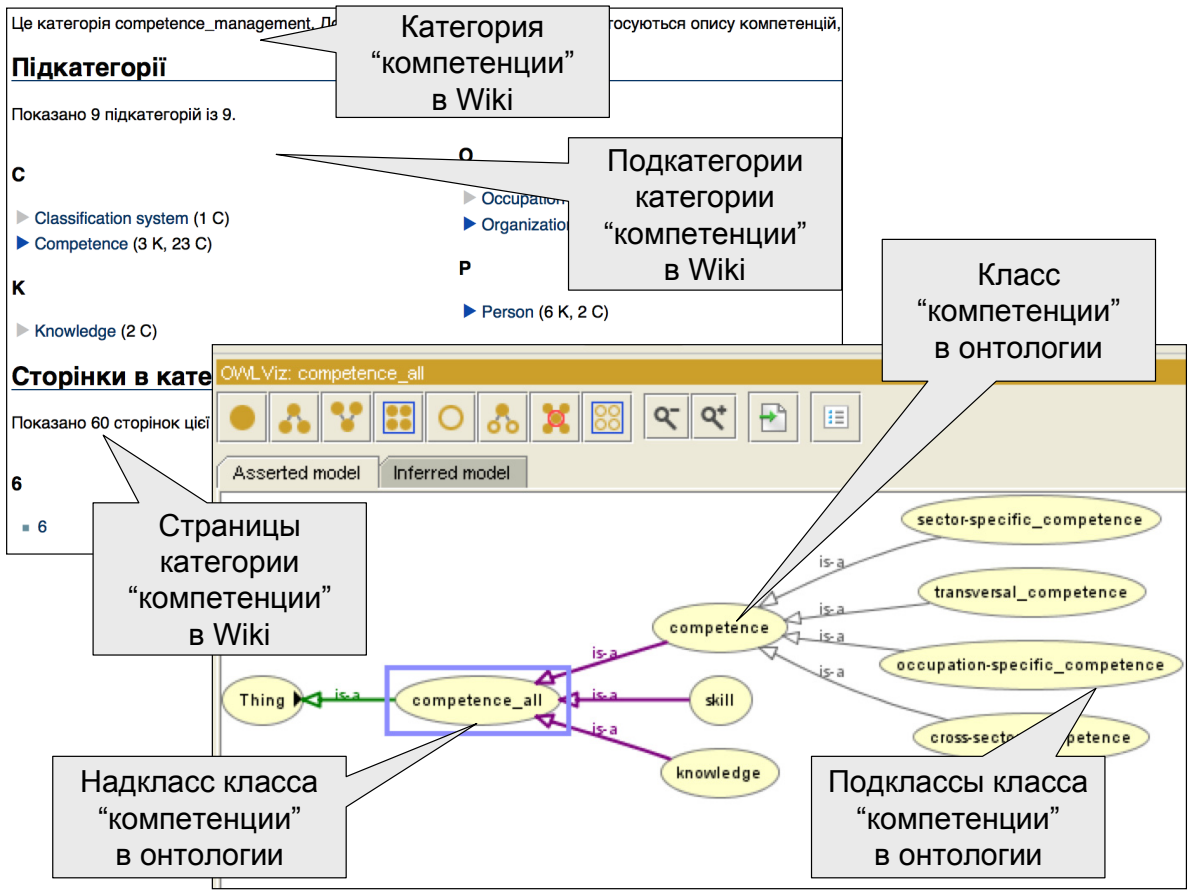

Рис.1. Соответствие между классом онтологии и категорией Wiki «Компетенция»

В общем случае для того, чтобы сопоставлять онтологии, нужно установить соответствие между их элементами. При сопоставлении версий онтологий задача значительно упрощается, так как большая часть соответствий между именами элементов устанавливается автоматически (имена не меняются), но необходимо найти возникшие отличия и оценить их влияние на Wiki-ресурс.

При этом важно, чтобы онтологическая модель соответствовала актуальному представлению Wiki-ресурса – иначе запросы не будут пертинентны потребностям пользователей (например, если семантическое свойство было переименовано, а в запросе используется старое имя, то значительная часть информации не будет обнаружена). Такая ситуация может возникнуть и в том случае, если разработчики Wiki-ресурса и его пользователи применяют разные версии онтологической модели. При этом возникает необходимость сопоставления версий онтологий.

Благодарности

В данном исследовании использованы результаты, полученные автором при выполнении госбюджетной темі III-2-12 “Разработка методов и средств поддержки построения интеллектуальных информационных систем в семантической Веб-среде” (Институт программных систем НАН Украины) и темы “Разработка интеллектуальной системы информационного и когнитивного сопровождения функционирования национальной рамки квалификаций” (Мелитопольский государственный педагогический университет).

Работа опубликована при частичной поддержке *ITHEA ISS (www.ithea.org) и ADUIS (www.aduis.com.ua).*

Выводы

Предложенный в работе подход к сопоставлению элементов онтологии с семантическими Wiki-ресурсами позволяет, с одной стороны, использовать преимущества онтологического представления информации для управления знаниями в интеллектуальных информационных системах (например, в различных Wiki-справочниках), а с другой, обеспечивает автоматизированный и простой для пользователей способ пополнения и усовершенствования онтологий различных предметных областей. Представляется целесообразным апробировать предложенные методы на более широком классе прикладных задач и обеспечить их соответствующими инструментами.

Литература

[Berners-Lee, 2001] Berners-Lee T., Hendle, J., & Lassil, O. The semantic web. Scientific american, 284(5), 28-37, 2001. – https://pdfs.semanticscholar.org/566c/1c6bd366b4c9e07fc37eb372771690d5ba31.pdf

[Davies, 2002] Davies J., Fensel D., van Harmelen F. Towards the Semantic Web: Ontology-driven knowledge management. – John Wiley & Sons Ltd, , England, 2002.

[Gladun, 2013] Gladun A., Rogushina J., Valencia-García R., Martínez-Béjar R. Semantics-driven modelling of user preferences for information retrieval in the biomedical domain. Informatics for health and social care. V.38, N.2, 2013. – P.150-170.

[Gruber, 1995] Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? .(). Toward principles for the design of ontologies used for knowledge sharing?. International journal of human-computer studies, 43(5-6), 907-928, 1995. – http://eolo.cps.unizar.es/docencia/doctorado/Articulos/Ontologias/Toward%20Principles%20for%20the%20Design%20of%20Ontologies%20Used%20for%20Knowledge%20Sharing-Gruber1995.pdf.

[Guarino, 1995] Guarino N. Formal ontology, conceptual analysis and knowledge representation. International journal of human-computer studies, 43(5-6), 625-640.1995. – https://www.researchgate.net/profile/Nicola_Guarino/publication/2368416_Formal_Ontology_Conceptual_Analysis_and_Knowledge_Representation/links/5593060d08ae16f493ee4d94.pdf.

[Guarino, 2000] Guarino N., Welty C. A formal ontology of properties. Proc. of International Conference on Knowledge Engineering and Knowledge Management, 97-112, 2000. – http://cuiwww.unige.ch/isi/cours/aftsi/articles/01-guarino00formal.pdf

[Holtshouse, 2013] Holtshouse, D. K. Information technology for knowledge management. U. M. Borghoff, & R. Pareschi (Eds.). Springer Science & Business Media, 2013. – http://tocs.ulb.tu-darmstadt.de/55883427.pdf

[Loukachevitch, 2014] Loukachevitch N., Dobrov, B. RuThes linguistic ontology vs. Russian wordnets. In Proc.of Global WordNet Conference GWC-2014, 2014. – http://www.aclweb.org/anthology/W/W14/W14-0121.pdf.

[Lundqvist, 2011] Lundqvist K. Ø., Baker K., Williams, S. (2011). Ontology supported competency system. International Journal of Knowledge and Learning, 7(3-4), 197-219. – http://centaur.reading.ac.uk/23841/1/authorFinalVersion.pdf.

[Majchrzak, 2006] Majchrzak A., Wagner C., Yates D. Corporate wiki users: results of a survey. In Proc.of the 2006 international symposium on Wikis, 99-104, 2006. – http://stu.hksyu.edu/~wkma/notes/jour395/wagner2004.pdf

[Mishra, 2010] Mishra R B, Sandeep K. Semantic Web Reasoners and Languages, Springer, 2010.

[Pedersen, 2004] Pedersen T., Patwardhan S., Michelizzi J. WordNet: Similarity - measuring the relatedness of concepts. Proc. Of the 19th National Conference on Artificial Intelligence (AAAI-04), 1024-1025, 2004.

[Rogushina, 2012] Rogushina J., Gladun A.Ontology-based competency analyses in new research domains. Journal of Computing and Information Technology. V.20, N. 4, 2012. – P.277-293.

[Rogushina, 2016] Rogushina J. Semantic Wiki resources and their use for the construction of personalised ontologies. Proc. of the 10th International Conference of Programming UkrPROG'2016, 196-203, 2016. – http://ceur-ws.org/Vol-1631/188-195.pdf

[Studer, 1998] Studer R., Benjamins V. R., Fensel D. Knowledge engineering: principles and methods. Data & knowledge engineering, 25(1-2), 161-197, 1998. – https://www.researchgate.net/profile/V_Richard_Benjamins/publication/222305044_Knowledge_engineering_principles_and_methods_Data_Knowl_Eng_251-2161-197/links/0fcfd50c3673c0368e000000/Knowledge-engineering-principles-and-methods-Data-Knowl-Eng-251-2161-197.pdf.

[Tudorache, 2013] Tudorache T., Nyulas C., Noy N. F., Musen M. A. (2013). WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web. Semantic web, 4(1), 89-99. – https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691821/.

[Uschold, 1996] Uschold M., Grüninger M. Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, 11(2), 93–155, 1996.

[Vargas-Vera, 2002] Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F. MnM: Ontology driven tool for semantic markup. Proc. Workshop Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002), 43-47, 2002. – http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.3590

[Völkel, 2006] Völkel M., Krötzsch M., Vrandecic D., Haller H., Studer, R. Semantic wikipedia. Proc.of the 15th international conference on World Wide Web.585-594, 2006. – http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.9471&rep=rep1&type=pdf.

[Гладун, 2016] Гладун А.Я., Рогушина Ю.В. Семантичні технології: принципи та практики. К.:ТОВ "ВД "АДЕФ-Україна", 2016. – https://core.ac.uk/download/pdf/38468940.pdf.

[Лесько, 2010] Лесько О.Н., Рогушина Ю.В. Использование специализированной лексической онтологии для автоматизации формирования онтологии предметной области по естественно-языковым текстам. Information Models of Knowledge. Edited by K.Markov, V.Velychko, O.Voloshin. – ITHEA, Kiev-Sofia, 93-100, 2010.

Информация об авторе

Рогушина Юлия *– к.ф.-м.н., доцент, с.н.с. Института программных систем НАН Украины; ORCID http://orcid.org/0000-0001-7958-2557, e-mail: ladamandraka2010@gmail.com.*

MATCHING OF SEMANTIC INFORMATION RESOURCES OF THE WEB ON THE BASIS OF ONTOLOGICAL ANALYSIS

Julia Rogushina

Abstract*: Modern means of representation of distributed knowledge are analyzed. The problems arising in the process of knowledge managing for intelligent information systems oriented to work on the Web are considered. The necessity of ontology use and the actuality of the development of semantic technologies aimed at their support, in particular, the standards and languages of the Semantic Web, are grounded. The problem of ontology matching, the results of such a matching and a particular case of this problem where compared ontologies are the development of the same initial ontology are considered. Semantic Wiki-resources are considered as a source of information for replenishment and improvement of ontologies of domain that is displayed by the set of Wiki pages. The analysis of components either of domain ontology or of semantic markup of Wiki-resources (on example of Semantic MediaWiki) is carried out, the method of their automated matching for the improvement of knowledge bases of distributed intellectual systems is proposed.*

Keywords*: domain ontology model, ontology, Wiki-technology, automatically generated ontologies, Subject Domain.*