УДК 519.7

Н. В. РЯБОВА

ОБ ОНТОЛОГИЧЕСКОМ ПОДХОДЕ К ПРЕДСТАВЛЕНИЮ ЛИНГВИСТИЧЕСКИХ ЗНАНИЙ В СИСТЕМАХ ИНТЕЛЛЕКТУАЛЬНОЙ ИНТЕГРАЦИИ ИНФОРМАЦИИ

В настоящее время всемирная информационная сеть Интернет предоставляет неограниченный доступ широкому кругу пользователей к огромным массивам самой разнообразной информации. Зачастую эта информация недостаточно хорошо структурирована, плохо организована и не очень корректно классифицирована. И разработчикам, и пользователям стало ясно, что Интернет, превратившийся в огромное информационное пространство, которым никто не управляет, скоро может уподобиться спутанному неводу, где рыба есть, но найти ее, а тем более достать, весьма проблематично. На повестку дня встал вопрос: "Как упорядочить работу в Интернет таким образом, чтобы поиск информации стал действительно смысловым и позволял получать релевантные запросу документы, как организовать взаимодействие различных распределенных информационных систем, которые должны предоставлять выборочную информацию из одного смыслового поля для дальнейшей ее обработки, обобщения, прогнозирования?".

Общеизвестно, что способы выражения знаний об одних и тех же вещах даже на одном и том же языке могут настолько различаться, что люди (не говоря уже о компьютерных системах) не понимают друг друга, даже при обсуждении одного и того же предмета. Это приводит к трудностям информационного обмена между людьми, организациями и программами. Особо актуальна эта проблема при разработке систем интеллектуальной интеграции информации, которые должны функционировать в открытых информационных пространствах типа Интернет и обеспечивать как можно более полный набор сервисных функций для ответа на запросы, требующие извлечения и комбинирования данных из множества Web-источников, включающих в себя гетерогенные и текстовые базы данных, где информация представляется, как правило, в разных форматах, с разной степенью детализации и на основе различных формализмов.

Необходимо устранить или свести к минимуму концептуальную и терминологическую путаницу и установить однозначное понимание языка, используемого для формирования требований и спецификаций сложных систем. Такой язык должен служить средством: 1) коммуникации между людьми, имеющими различный взгляд на одни и те же вещи; 2) взаимодействия между программными системами путем трансляции в него и из него; 3) инструментальной поддержки для повторного использования благодаря формальной спецификации, унификации представления различных моделей, автоматизации проверки корректности, переводимости различных методов моделирования в унифицированное представление [1].

Для решения этих и многих других проблем, связанных с семантической обработкой информации в Web, коллективы разработчиков в области искусственного интеллекта (ИИ) выдвинули лозунг: "Превратить информационное пространство Интернет в пространство знаний". Можно с уверенностью сказать, что сегодня все крупнейшие разработки в области искусственного интеллекта направлены на решение этой глобальной проблемы.

Известны такие проекты, как CYC [2, 3], WordNet [4, 5], Generalized Upper Model [6, 7], SemanticWeb [8], которые уже добились весомых результатов в этом направлении, однако большая часть работы еще не сделана. Объясняется это многими причинами. World Wide Web является информационным ресурсом с виртуально неограниченным потенциалом. Тем не менее этот потенциал не используется в должной степени из-за нерешенной проблемы автоматического извлечения смысла из содержащейся в Web информации (прежде всего, в виде естественноязыковых текстов – ЕЯТ), ее последующей обработки и интеграции. Таким образом, проблема Semantic Web тесно коррелирует с исследованиями в области извлечения знаний из ЕЯТ и, следовательно, требует от разработчиков описания документов с использованием языков представления знаний.

Однако, проблема представления знаний в Web имеет свои особенности. В отличие от большинства традиционных баз знаний Web характеризуется высокой степенью децентрализации, а также частыми изменениями в огромных информационных массивах. Следовательно, необходимы такие подходы к представлению знаний, которые могли бы учитывать эти факторы и адаптироваться к ним. На сегодняшний день самым перспективным подходом к представлению знаний в WWW признан *онтологический подход*, позволяющий интегрировать информацию из Web-источников на базе онтологий. При этом взаимодействующие источники информации должны нести так называемые

"онтологические обязательства" (ontological commitment), т.е. использовать общую систему понятий и отношений между ними, описывающих концептуальную модель предметной области и организованных в таксономию.

Термин "онтология" заимствован разработчиками систем искусственного интеллекта из философии, где он обозначает "учение о бытии, о сущем, в отличие от гносеологии – учения о познании" [9]. Онтология как научная дисциплина изучает различные типы абстрактных и общих философских сущностей и категорий. Исследователи в области ИИ перенесли этот термин в свою область и зачастую используют его для названия документа или файла, в котором формально определены отношения на множестве применяемых в изучаемой области терминов. Наиболее типичная разновидность онтологии, используемой в Web, имеет таксономию и множество правил вывода. В самом общем случае можно сказать, что онтология – это представленные на некотором языке, обладающем свойствами, перечисленными в [1], знания о некоторой области интересов (среде, мире). Поскольку направление ИИ, связанное с представлением знаний (прежде всего в Интернет-пространстве) на базе онтологического подхода и разработки онтологий, является сравнительно новым, до сих пор продолжаются дискуссии по поводу терминологии, методологии и различных сфер применения онтологий. Тем не менее из области инженерии знаний (*Knowledge Engineering*) прочно выделилось и активно развивается направление, получившее название "онтологический инжиниринг" (*Ontological Engineering*). В рамках этого подхода все классические методы и модели представления знаний в ИИ получили как бы второе дыхание и активно адаптируются разработчиками для функционирования в новой виртуальной информационной среде. Предметом исследований онтологического инжиниринга является изучение категорий сущностей, объектов, предметов, которые существуют или могут существовать в некоторой предметной области. Результатом такого исследования должна стать онтология, представляющая собой каталог классов или типов сущностей, которые должны существовать в исследуемой области D с точки зрения исследователя, использующего язык L для описания данной области. Типы в онтологии представляют предикаты, смыслы слов (понятия) или концепты, и типы отношений, выраженные средствами языка L. Все эти средства используются для обсуждения различных тем (т. е. кооперативного решения проблем) в области D. Неинтерпретируемая логика (например, исчисление предикатов первого порядка, концептуальные графы) является онтологически нейтральной. Она не накладывает никаких ограничений на предмет обсуждения или способ выявления его основных характеристик. Логика сама по себе не может ничего сказать о каком-либо предмете или области исследований, но комбинация логики с онтологией дает язык, который может выражать связи и отношения между сущностями в исследуемой области.

Термин *онтология* широко используется разработчиками современных Интернет-систем как аналог понятия модели какой-либо конкретной предметной области и охватывает все необходимые знания, формализованные, как правило, в виде некоторой формальной теории, и интерпретатор или машину вывода. Таким образом, можно сказать, что, несмотря на некоторую терминологическую путаницу и продолжающиеся дискуссии, онтологический подход к представлению знаний в открытых сетевых пространствах набирает силу и имеет самые широкие перспективы.

Общепризнанным в среде разработчиков ИИ является определение Тома Грубера [10]: "Онтология – есть эксплицитная спецификация концептуализации". В этом определении концептуализация понимается как абстракция определенного взгляда на мир в соответствии с определенным образом действий и может быть представлена в виде кортежа $\langle D, R \rangle$, где D есть область дискурса (рассуждения); R – множество отношений на D. Онтология соединяет термины словаря с категориями, выделенными при концептуализации, и предусматривает определения, накладывающие ограничения на области интерпретации этих терминов. Однако такое определение концептуализации не может быть использовано в определении онтологии, поскольку здесь предполагается, что концептуализация представляет собой единственное константное состояние внешнего мира, т. е. является экстенсиональной, константной структурой. Между тем онтология должна быть снабжена всеми терминами, необходимыми для представления описания всех возможных состояний рассматриваемой предметной области при изменениях внешней среды. В связи с этим концептуализацию целесообразно представлять в виде интенсиональной структуры [12] $\langle W, D, R \rangle$, где W – множество возможных миров; D – область дискурса; R – множество всех возможных n-арных интенсиональных отношений. В общем случае онтология – это спецификация концептуализации, которая состоит из словаря и теории. Онтологии включают в себя абстрактное описание

очень общих, так и специфичных для конкретной предметной области терминов. Но кроме спецификации семантики выделенного в онтологию множества терминов, онтологии обладают такими сильными потенциальными свойствами, как разделение знаний и их повторное использование, а также расширяемость и ревизия (проверка и модификация при необходимости) одних онтологий другими.

Различают следующие типы онтологий [11]. Неформальная онтология (*informal ontology*) может быть специфицирована каталогом типов или классов, которые, в свою очередь, могут иметь или не иметь определений в виде утверждений на естественном языке. Формальная онтология (*formal ontology*) специфицируется набором имен для концептов и типами отношений, частично упорядоченных с помощью отношения "тип–подтип". В формальной онтологии достигается большая строгость в установлении различий между классами, поскольку используются понятия "подтип" и "супертип". При этом в аксиоматизируемой онтологии (*axiomatized ontology*) подтипы различаются с помощью аксиом и определений на формальном языке, например, на языке логики предикатов или некоторой компьютерно-ориентированной нотации, которая может быть транслирована в логический язык. Онтология, основанная на прототипах (*prototype-based ontology*), различает подтипы, сравнивая их с типичным образцом каждого прототипа.

Сегодня перечень проектов, которые в той или иной степени связаны с онтологиями для Web, уходит за горизонт, среди них весомый вклад вносят лингвистические онтологии. К числу наиболее известных проектов разработки лингвистических онтологий относится CYC – создание мультиконтекстной базы знаний и машины вывода, разрабатываемой Cycorp. Основная цель этого гигантского проекта – раз и навсегда построить базу знаний всех общих понятий, включающую в себя семантическую структуру терминов, связей между ними, правил, которая будет доступна разнообразным программным средствам. Проект CYC, начатый в 1984 г., является смелой попыткой сконструировать основу для представления наиболее общих знаний и реализации вывода на этой системе знаний. Проект продолжается и по сегодняшний день. CYC можно использовать как стандартную онтологию верхнего уровня (top level), работающую с самыми необходимыми, общими понятиями (включающую словарь концептов и организационную структуру) и служащую основой для создания онтологий прикладных областей (например, в электронной коммерции) и базовой основой для стандартизации при создании более эффективных систем информационного поиска, интеграции информациии, управления знаниями и т. д. Прикладные прототипы всех этих задач уже работают [3].

Проект WordNet [4,5], разработанный в Cognitive Science Laboratory Принстонского университета под руководством профессора Д. Миллера, является на сегодняшний день одной из наиболее известных, обширных и обстоятельных лексических онтологий, хотя скромно определяется авторами как онлайновая лексическая справочная система, проектирование которой основывалось на современных психолингвистических теориях организации лексической памяти человека. В систему входят английские имена существительные, глаголы, прилагательные и наречия, организованные в множества синонимов, каждое из которых представляет один основной лексический концепт. Связи между множествами синонимов устанавливаются с помощью различных отношений. Лексические объекты в WordNet организованы по семантическому принципу (с базовым разделением на имена существительные, глаголы, прилагательные и наречия). Центральным объектом в WordNet является множество синонимов, называемое “синтетическим множеством” (*synset*). Если слово имеет более одного смысла, оно появляется в нескольких множествах *synset*. Всего в системе более 70000 таких множеств, организованных в иерархические структуры с помощью отношения суперкласс-подкласс (соотносимого с лингвистическим отношением гипернимии–гипонимии, hypernymy-hyponymy). Для каждого концепта (представленного соответствующим множеством *synset*), устанавливаются указатели-поинтеры (pointers) к существительным, представляющим части данного концепта. Например, для концепта "птица" (bird) установлены поинтеры "клюв" (beak) и "крылья" (wings). Кроме того, в WordNet используются и другие типы поинтеров (например, поинтеры от существительных к глаголам реализуют отношения типа *функции*, а от существительных к прилагательным – отношения типа *свойства*). По своей сути WordNet представляет собой таксономию и не имеет структурированных концептов или аксиом. Хотя WordNet использует простую иерархию для множеств *synset* на корпусе имен существительных английского языка, в то же время он предусматривает различную организацию *synset*-множеств для глаголов и прилагательных. Качественные прилагательные, описывающие какие-либо свойства (*descriptive*

adjectives), организованы в биполярные кластеры по принципу антонимии. Например, биполярный кластер образуется из прилагательных "сухой" (*dry*) и "мокрый" (*wet*), где каждое из них представляет отдельное *synset*-множество, содержащее синонимы соответствующего концепта. Такие относительные прилагательные, как например, *fraternal* (братский) в *fraternal twins* (двуяйцевые близнецы), организованы только в *synset*-множества с поинтерами к соответствующим существительным. Глаголы в WordNet разбиты на 15 кластеров в соответствии с их значениями и установлением основного, базового значения между глаголами в кластере. Большинство этих кластеров соответствует семантическим областям разбиения, например: 1) *bodily functions and care* (глаголы с общим значением физических функций и обслуживания, ухода); 2) *change* (один из самых больших файлов в WordNet, включает около 750 *synset*-множеств, область делится на 6 подобластей с общими значениями изменения, изменения состояния, изменения в обратную сторону, возвращения, полного изменения, превращения, изменения формы, адаптации); 3) *cognition* (познание, узнавание, распознавание), *communication* (глаголы, передающие значения коммуникации, процессов передачи информации), 4) *competition* (глаголы с общим значением конкурирования, соперничества). Часть глаголов, такие как, *suffice* (удовлетворять, быть достаточным для чего-либо), *belong* (подходить, соответствовать; принадлежать; быть частью), *resemble* (иметь сходство, походить), не принадлежащие ни к одной из семантических областей, образуют отдельный файл. Заметим, что полный анализ гигантского проекта WordNet невозможен в рамках одной статьи, и хотя работа над этой системой продолжается, многие исследователи уже сейчас успешно используют его как лингвистическую онтологию для решения таких задач, как усовершенствование поиска естественноязыковых текстов, создание систем содержательного, смыслового поиска информации в электронных каталогах и желтых страницах Интернет [12-14], при проектировании электронных библиотек и др.

Generalized Upper Model [6,7] представляет собой "лингвистически-мотивированную" предметно-независимую онтологию для решения наиболее общих задачах, которая поддерживает усовершенствованную, усложненную обработку естественно-языковой информации в различных языках (английский, немецкий, итальянский) и в то же время, по утверждению разработчиков, значительно упрощает интерфейс между специфическими знаниями конкретных предметных областей и основными лингвистическими ресурсами. Предполагается также, что этот проект может послужить хорошей основой для моделирования предметных областей (например, при проектировании информационных корпоративных систем или систем управления знаниями). Уровень абстракции в данной онтологии является промежуточным между лексическими и концептуальными знаниями.

Попытки структурировать Web предпринимаются постоянно. Онтологии – это еще одна попытка решить проблему информационного переполнения в сети. В настоящее время широко обсуждается проблема: какими быть онтологиям в Интернет – универсальными или ограниченными предметной областью (областями); уникальными в системе или состоящими из распределенного подмножества; доступными для редактирования всем пользователям или только администратору. Актуальными направлениями исследований являются: поиск новых структурных решений внутренней организации онтологии, способов доступа к хранящейся в ней информации, новых методов вывода и представления информации для пользователя. Кроме того, необходимо развивать методы взаимодействия пользователя с онтологией (онтологиями). Основные задачи, которые могут успешно решаться (и решаются) на базе онтологий, включают в себя предоставление знаний для вывода информации, которая релевантна запросу пользователя; фильтрацию и классификацию информации; индексирование собранной информации; организацию общей терминологии, которой могут пользоваться для коммуникации программные агенты и пользователи. До сих пор возможности логического вывода в Интернет практически не применялись. С "приходом" баз знаний и систем, основанных на знаниях, в Web появляются новые перспективы в освоении сетевого пространства.

Лингвистические онтологии представляют собой огромный потенциал для сбора, комплектования, накопления информации из Web-ресурсов. Даже судя по их сегодняшнему состоянию (с бедными онтологическими структурами), они могут обеспечить существенное улучшение работы современных поисковых систем. Преобразование их в доступные, ясные, обогащенные и логически связные, согласованные онтологии, применимые для управления информационно-поисковыми системами, не является легкой задачей, но ценность таких онтологий будет постоянно возрастать в связи с постоянным расширением Интернет и обвальным ростом

естественноязыковой информации, представленной в нем в виде различных документов, текстовых баз данных, электронных каталогов, "желтых страниц" и т. п.

Список литературы: 1. *Девятков В.В.* Онтологии в проектировании систем. http://inftech.webservis.ru/it/conference/scm/1999/devyatkov.html. 2. *Lenat D.B.* CYC: A Large-Scale Investment in Knowledge Infrastructure // Communications of the ACM. 1995. Vol. 38, №11. P.33-38. 3. http://www.cyc.com/cyc-2-1/cover.html. 4. *Miller G.A.* WORDNET: An On-Line Lexical Database // International Journal of Lexicography. 1990. №3-4. P.235-312. 5. ftp://clarity.princeton.edu/pub/wordnet/. 6. *Bateman J.A., Magnini B., Rinaldi F.* The Generalized Italian, German, English Upper Model // Proceeding of Eleventh European Conference on Artificial Intelligence (ECAI'94). Workshop on Comparision of Implemented Ontologies. 1994. Amsterdam, The Netherlands. 7. http://www.darmstadt.gmd.de/publish/ komet/genum/newUM.html. 8. *Martin P., Eklund P.* Embedding Knowledge in Web Documents // Proceedings of Eight International World Wide Web Conference. Toronto. 1999. 9. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб: Питер, 2000. 10. *Gruber T.R.* A Translation Approach to Portable Ontologies // Knowledge Acquisition. 1993. N5 (2). P.199-220. 11. *Sowa J.F.* Knowledge Representation: Logical, Philosophical and Computational Foundations. Pacific Grove: Brooks Cole Publishing, 2000. 594 p. 12. *Gandemi A., Guario N.,. et. al.* Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level // LADSEB-CNR Internal Report. 2001. N 6 http://saussure.irmkant.rm.cnr.it . 13. *Gonzalo J., Verdejo F., Chugar I., Cigarran J.* Indexing with WordNet synsets can improve text retrieval // http://www.lanl.gov.cmp-lg/9808002. 14. *Guarino N., Masolo K., Vetere G.* OntoSeek: Content-Based Access to the Web // IEEE Intelligent Systems. 1999. N 14(3). P.70-80.

Поступила в редколлегию 12.11.2001г.