УДК 004.89

СЕМАНТИЧЕСКАЯ ПАУТИНА И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

И.А. Бессмертный

В статье рассматривается проблема создания искусственного интеллекта, моделирующего человеческое мышление. Предлагается способ количественной оценки естественного и искусственного интеллекта и оценивается необходимый объем базы знаний. Рассматривается возможность создания глобального искусственного интеллекта путем использования в качестве базы знаний Семантической паутины – Всемирной паутины нового поколения. Анализируются проблемы практической реализации данной идеи и возможные пути решения.

Ключевые слова: искусственный интеллект, Семантическая паутина, интеллектуальный агент.

Введение

Появление машин и механизмов привело к тому, что физическая сила и выносливость перестали быть основными свойствами человека как работника. Уместно также вспомнить «Великий уравнитель» – кольт. В настоящее время мы наблюдаем, как благодаря Интернету утрачивает ценность эрудиция человека, поскольку практически любая информация может быть доставлена пользователю в считанные секунды. Однако Интернет дает только доступ к документам, оставляя человеку извлечение и интерпретацию данных. Логичным развитием Интернета является объявленный в концепции Семантической паутины (Semantic Web, СП) переход от извлечения документов к извлечению знаний, а также к их автоматической обработке [1]. База знаний, образуемая множеством семантических документов СП, вместе с интеллектуальным агентом (ИА) для извлечения знаний образуют структуру, напоминающую искусственный интеллект [2]. В статье рассматривается вопрос, может ли СП стать средой для создания глобального искусственного интеллекта (ИИ), позволяющего не только предоставлять пользователю факты, но и выполнять умозаключения и порождать новые знания.

Состояние проблемы и текущие исследования

Работы по практическому созданию ИИ велись до появления Интернета. Наиболее масштабная разработка принадлежит Дагласу Ленату, основателю компании Cycorp (www.cyc.com), который, начиная с 1984 г., формализовал и тщательно классифицировал несколько миллионов единиц знаний на бытовые темы. Однако, как замечает известный футуролог и эксперт в области ИИ Марвин Мински, этого все еще недостаточно для того, чтобы отвечать на вопросы, доступные трехлетнему ребенку [3]. В рамках концепции СП в настоящее время ведется активная работа по созданию онтологий в качестве первого этапа масштабной формализации знаний. В частности, в Стенфордском университете разработана программная платформа – редактор онтологий Protégé (protege.stanford.edu), а также организовано сообщество энтузиастов, насчитывающее несколько тысяч участников, которые пополняют базу онтологий для самых разных предметных областей. Заслуживает внимания также проект Estrella (www.estrellaproject.org/lkif-core), в рамках которого разработана онтология LKIF (Legal Knowledge Interchange Format) – язык для представления юридических знаний и обмена между правовыми информационными базами. Среди отечественных разработок следует отметить онторедактор InTez, создаваемый в СПбГУ [4].

По мнению Г. Осипова [5], основным препятствием на пути создания ИИ является отсутствие алгоритмов и начальной компетентности. Поскольку ИА решает классическую задачу поиска на дереве решений, основная проблема заключается в ее размерности, так как по мере увеличения числа правил задача поиска очень быстро приобретает астрономические масштабы. Отсутствие начальной компетентности хорошо иллюстрируется опытом вышеупомянутого Д. Лената, у которого компьютер все еще нуждается, чтобы ему объясняли, что родители старше детей, а люди перестают выписывать газеты, если умирают. Современный Интернет также страдает от недостатка информации начального уровня. Убедиться в этом легко, вооружившись компьютером во время просмотра телепрограммы «Кто хочет стать миллионером». Интернет позволяет легко найти ответы на сложные вопросы, в то время как самые простые вопросы, скорее всего, останутся без ответа. Следует отметить, что формализация знаний является сложным творческим процессом, темп которого соизмерим со скоростью приобретением знаний.

В этой связи целесообразно оценить, какой объем знаний должна содержать база знаний ИИ и в каких единицах этот объем оценивать.

Количественная оценка интеллекта

Проблема оценки ИИ была поднята задолго до появления возможностей его реализации. Классическим примером является тест Тьюринга [2], который предполагает качественную оценку ИИ в сравнении с естественным интеллектом (ЕИ). Между тем хотелось бы располагать количественной мерой объема базы знаний, которыми владеет человек и которыми должен оперировать ИИ.

В первую очередь следует определить единицы измерения знаний, одинаково пригодные для оценки как ЕИ, так и ИИ. Универсальной и широко используемой единицей оценки объема приобретенных знаний человеком является время, затраченное на обучение. Другой подход – символьный, исходящий из количества бит данных, находящихся в учебниках и других источниках, используемых для обучения. И первый, и второй способы отражают лишь потенциал для приобретения знаний, но не результат обучения. Очевидно также, что эти меры не могут применяться к оценке ИИ. По нашему мнению, заслуживает внимания методика, предложенная И.В. Богдановым [6], в которой в качестве единицы измерения количества теоретических знаний принято ис-

пользовать понятие как отражение в сознании человека общих и существенных свойств предметов и явлений в мыслеобразе определенного объема. За единицу измерения количества практических навыков и умений принимается умение, определяемое как законченное действие, состоящее из некоторых операций. Объем умения определяется числом действий, из которых оно состоит. Данная методика хорошо накладывается на элементы ИИ: понятие соответствует сущности (классу или экземпляру класса), а умение – правилу в базе знаний.

Применяя описанную выше методику, можно попытаться приблизительно оценить объем ЕИ. Для этого можно использовать несколько способов.

Первый способ – это преобразование времени обучения в объем знаний. В работе [8] приводятся результаты исследований, согласно которым за один академический час студенты усваивают от 30 до 50 понятий, каждое из которых в среднем содержит три связи. Если применить это значение темпа усвоения знаний к суммарному фонду времени обучения в средней школе и вузе, то получим 500–900 тыс. понятий или умений. По результатам исследований [7], в активной памяти остается 3–10% усвоенных знаний. Следовательно, эти цифры следует уменьшить по меньшей мере на порядок. С другой стороны, человек приобретает знания и опыт непрерывно, за исключением времени сна. Если применить и к этому процессу вышеуказанный темп усвоения знаний, 30–50 понятий в час (маловероятно, что стихийное усвоение знаний более эффективно, чем занятия с педагогами), то можно получить верхнюю оценку совокупных знаний 20-летнего индивидуума в пределах 4–6 миллионов понятий. Реалистичная оценка активных знаний с учетом фактора усвоения должна лежать в пределах 500 000 понятий и умений, из которых превалирующую часть составляют вовсе не научные, а элементарные бытовые знания и умения.

Статья	Предложений	Статья	Предложений
Alaska Route 10	3	Six Degrees	46
BSFOCS (Black Sea Fibre Optic Cable System)	3	Incubation: Time Is Running Out	18
Ces Gens-Là (song)	7	House Finch	46
D. V. Gundappa	30	Calends	19
es (operating system)	4	Henry William Bristow	5
F. E. Compton	5	Craig Clark	3
Catalan Sheepdog	20	Bekka Bramlett	5
Drug Enforcement Administration	101	Archibald Higgins	2
Phong Nha-Ke Bang National Park	285	Federal Water Power Act of 1920	4
Sé (Funchal)	4	Houron	3
Recognizance	5	The Mall at Tuttle Crossing	10
Franz Rosenthal	20	Jeffersonville, Kentucky	18
Short Admiralty Type 74	11	Battle of Cape Spartel	33
Henry P. Deuscher House (Ohio)	1	Jasem Yacob	2
Hibiscadelphus hualalaiensis	3	Avex Group	26
Среднее число предложений		25	

Таблица. Количество предложений в статьях Википедии

Второй способ – оценка объема знаний через энциклопедии и словари. Последняя англоязычная редакция энциклопедии «Британника» содержит 120 000 статей, «Большая Советская Энциклопедия» – около 100 000 статей, «Толковый словарь живого великорусского языка» В. Даля – 44 694 статьи, «Русский орфографический словарь» – около 180 000 слов. Самым богатым языком считается английский, который насчитывает 490 000 слов и еще 300 000 технических терминов [9]. В словарях, в частности, в

Русском орфографическом словаре, отсутствуют понятия, описываемые более чем одним словом, зато однословные понятия дублируются за счет синонимов. Особого внимания заслуживает Википедия как плод коллективного творчества многих тысяч авторов. В данный момент англоязычная версия Википедии насчитывает свыше 3 млн. статей, что существенно больше, чем в традиционных энциклопедиях, и объясняется ее гипертекстовой организаций, позволяющей выделять в отдельные статьи частные понятия, не удостоенные отдельных статей в печатных изданиях. Приблизительно оценить количество знаний в энциклопедиях можно, подсчитав количество предложений в одной статье в предположении, что одно предложение соответствует одному понятию. Случайная выборка 30 статей из англоязычной версии Википедии (см. таблицу) показала, что в среднем одна статья содержит около 25 предложений. Таким образом, вся Википедия хранит около 70 млн. понятий, причем это количество соответствует не индивидуальному интеллекту, а интеллекту всего человечества, и эта величина приблизительно в 140 раз больше объема знаний индивидуума.

Семантическая паутина как база знаний искусственного интеллекта

Полученная выше оценка объема знаний ЕИ в значительной степени объясняет более чем скромные успехи в соревновании ИИ и ЕИ [3]. Создание ИИ требует миллионов часов времени специалистов как в области инженерии знаний, так и в каждой предметной области. Это под силу лишь многотысячному высокоорганизованному коллективу, а финансирование такой разработки потребует сотен миллионов долларов; при этом практическая отдача может заключаться всего лишь в успешном прохождении теста Тьюринга. В таких условиях концепция СП появилась как нельзя кстати. Во-первых, это среда, изначально предназначенная для хранения знаний. Во-вторых, идеология СП предусматривает возможность участия в пополнении базы знаний неограниченного множества авторов. Опыт Википедии показывает, что такое сотрудничество возможно, причем на добровольных началах и с минимальным вмешательством координаторов проекта. В-третьих, кроме большого числа авторов, возможно привлечение множества тестировщиков баз знаний. В-четвертых, распределенное хранение знаний позволяет обойтись без мощных серверов.

Таким образом, концептуально СП может рассматриваться в качестве носителя базы знаний для ИИ. Попытаемся оценить, насколько эта идея реализуема на практике. Для извлечения и логической обработки знаний (reasoning) требуется специальная программа – интеллектуальный агент. Концепция СП, помимо методов представления знаний, предусматривает также и моделирование рассуждений с помощью ИА. Попытки создания ИАСП на данный момент представлены, в основном, немногочисленными экспериментальными разработками. В качестве примера можно привести винный агент, созданный в лаборатории ИИ Стенфордского университета (http://onto.stanford.edu:8080/wino/index.jsp). Создание универсального ИА, на первый взгляд, не представляет проблем. Машина вывода (inference engine), используемая в экспертных системах, а также в интерпретаторе языка Prolog, может унифицировать цель с фактами и правилами в базе знаний, т.е. реализовать механизм обратного логического вывода. Столь же просто может быть реализован прямой логический вывод от известных фактов. Проще говоря, ИИ в среде СП можно представить как большую ЭС, оперирующую распределенной базой знаний.

Семантическая паутина и экспертные системы

В принципах построения ЭС и СП много общего. Как ЭС, так и СП используют базу знаний, состоящую из фактов и правил. Эквивалентом машины логического выво-

да ЭС в СП выступает интеллектуальный агент (ИАСП). Используя исходные данные пользователя, машина вывода ЭС и ИАСП решают поставленную задачу (резолюцию цели). Однако на этом сходство заканчивается. Главное отличие СП от ЭС заключается в том, что это инструменты для решения разных задач. Если целью СП является нахождение знаний, то ЭС предназначены для извлечения навыков, т.е. решения практических задач, основными из которых являются задачи классификации и конструирования. Иными словами, база знаний ЭС в основном содержит алгоритмы, овеществляющие опыт экспертов и позволяющие находить кратчайшие пути к цели. Назначение ЭС определяет их свойства – узкую специализацию каждой базы знаний и, следовательно, сильную зависимость от контекста.

База знаний СП может содержать знания, достаточные для нахождения решения, аналогичного результатам работы ЭС, но размерность задачи поиска может оказаться неоправданно большой. К сожалению, увеличение вычислительной мощности здесь не даст результата, поскольку поиск на дереве решений неизбежно потребует выявления дополнительных фактов в диалоге с пользователем. Например, попытка решить задачу медицинской диагностики приведет к запросу большого числа анализов, не относящихся к диагнозу, но лежащих на дереве решения, обход которого осуществляется в произвольном порядке. По этой причине большинство людей не лечатся по медицинской энциклопедии, а идут к врачу. База знаний ЭС обычно хранит отдельное дерево решений для каждой проверяемой гипотезы, чем и обусловливается высокая скорость нахождения решений при наличии сотен и тысяч правил.

Основное достоинство ЭС – способность быстро находить известные решения – одновременно является их недостатком. ЭС не могут найти решение, ранее не описанное экспертом. Базы знаний ЭС содержат большое количество фактов и правил предметной области, но в них обычно отсутствуют общие знания, объем которых на порядки превышает объемы специальных знаний [10]. Отсутствие таких знаний не дает машине вывода ЭС возможности установить необходимые причинно-следственные связи. Обширная база знаний СП дает теоретическую возможность находить ранее неизвестные решения, но даже известные решения каждый раз должны отыскиваться с нуля.

Проблема комбинаторной сложности и пути ее решения

Резолюция правил является обычной задачей неинформированного поиска, и в случае решения такой задачи методом простого перебора («наивный» поиск) количество развертываемых вершин дерева решений при поиске «сначала в ширину» $N = b^d$, где b – коэффициент ветвления, d – глубина самого поверхностного решения [2]. Пусть база знаний содержит n фактов и r правил, в теле каждого из которых имеется c условий. Тогда коэффициент ветвления (число попыток применения правил ко всем фактам) равен

$b = r\, n^c$.

Уже эта величина является достаточно большой: для базы знаний из 100 фактов и 10 правил по 3 условия в каждом $b = 10 \times 100^3 = 10^7$, что потребует нескольких часов вычислений. Если же решение должно получиться путем вывода из цепочки правил длиной d, то время, затрачиваемое на развертывание N вершин, станет совершенно неприемлемым.

Сокращение размерности задачи поиска возможно путем декомпозиции базы знаний, как, например, в проекте Cyc, развиваемом вышеупомянутой компанией Cycorp, Inc., где знания отдельных предметных областей группируются в микротеории. Иерархия микротеорий теоретически позволяет ограничиваться предметными областями, лежащими на одной ветви дерева. Однако, как показано выше, даже маленькие базы знаний могут порождать очень разветвленное дерево поиска.

Существенное ускорение резолюции правил обеспечивает алгоритм Rete [11], суть которого заключается в том, что для каждого правила строится префиксное дерево, узлы которого хранят факты, соответствующие условиям. Скорость здесь достигается за счет памяти (многократного дублирования фактов в вершинах деревьев). Уязвимым местом алгоритма Rete является необходимость обновления всего префиксного дерева при каждом добавлении или изменения факта.

Другой подход к ускорению извлечения знаний – это вывод на основе прецедентов (CBR, case-based reasoning) [6]. Применительно к базам знаний СП это означает применение к фактам всех возможных правил и добавление новых фактов к уже существующим.

Автор разработал и опробовал еще один подход к ускорению вывода – индексацию фактов. При чтении из базы знаний фактов в виде субъект–предикат–объект к ним добавляется уникальный номер в пределах данной базы, после чего строится индекс в виде «терм, роль, список номеров фактов», где терм – идентификатор, используемый в фактах, роль – место его использования (в качестве субъекта, предиката или объекта). Когда возникает необходимость применения правила, с помощью индекса выполняется отбор только тех фактов, которые содержат термы, используемые в условиях. После этого с помощью операций пересечения и объединения множеств фактов находится релевантное подмножество, которое и используется в резолюции правила. В результате достигается ускорение вывода не менее чем на два порядка. Основное преимущество данного метода перед алгоритмом Rete заключается в том, что не требуется построение громоздкого префиксного дерева, а индекс формируется во время чтения базы знаний и практически не отнимает дополнительного времени.

Для исследования методов построения баз знаний на основе семантических сетей и интеллектуальных агентов автор разработал программу Semantic [12], которая поддерживает создание баз знаний с фактами и правилами, аналогичными СП, но легко читаемыми человеком, графическую визуализацию и извлечение знаний путем развертывания графов либо с использованием примитивного подмножества естественного языка. Последняя версия программы написана на языке Visual Prolog 7.2 и поддерживает прямой вывод с индексацией фактов, а также создание и сохранение прецедентов.

Заключение

Концепция СП может служить основой для создания глобального искусственного интеллекта. Основным препятствием для реализации интеллектуального агента СП является комбинаторная сложность задачи поиска. В настоящее время на кафедре вычислительной техники СПбГУ ИТМО проходит апробацию программная среда для изучения основ и исследования способов организации баз знаний на принципах СП с использованием индексации фактов и механизма прецедентов.

Литература

1. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American Magazine. – May, 2001.
2. Рассел С., Норвиг П. Искусственный интеллект: Современный подход. – 2-е изд. / пер. с англ. – М.: Изд. дом «Вильямс», 2006.
3. 2001's Computer as Dream and Reality. The Discover Interview: Marvin Minsky [Электронный ресурс]. – Режим доступа: http://discovermagazine.com/2007/jan/interview-minsky/, свободный.

4. Рубашкин В., Пивоварова Л. Онторедактор как комплексный инструмент онтологической инженерии // Материалы межд. конф. «Диалог-2008», 2008.
5. Осипов Г. Искусственный интеллект: Состояние исследований и взгляд в будущее [Электронный ресурс]. – Режим доступа: http://www.raai.org/about/persons/osipov/pages/ai/ai.html, свободный.
6. Богданов И.В. Учебная информация и единицы ее измерения // Труды СГУ. – Вып.44. Гуманитарные науки. Психология и социология образования. – М.: СГУ, 2002.
7. Чмыхова Е.В., Богданов И.В. Особенности формирования объема знаний в виртуально-тренинговой технологии модульного обучения // Труды СГУ. – Вып.44. Гуманитарные науки. Психология и социология образования. – М.: СГУ, 2002.
8. Книга рекордов Гиннесса. – М.: Прогресс, 1991.
9. Wood L. Cycorp: The Cost of Common Sense // Technology Review. – March, 2005.
10. Forgy C.L. RETE: A fast algorithm for the many pattern / many object pattern match problem // Artificial Intelligence. – 1982. – V. 19. – P. 17–37.
11. Bessmertny I., Kulagin V. Semantic Network as a Knowledge Base in Training Systems // Proceedings of 11th IACEE World Conference on Continuing Engineering Education. Atlanta, GE, USA, 2008. – P. 95–99.

Бессмертный Игорь Александрович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, кандидат технических наук, доцент, igor_bessmertny@hotmail.com