РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2007, том 43, № 6

УДК 004.382

ИНСТРУМЕНТАЛЬНАЯ СРЕДА ИНТЕГРАЦИИ ПРОФЕССИОНАЛЬНЫХ ЗНАНИЙ, ПРЕДСТАВЛЕННЫХ В ВИДЕ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

И. А. Минаков

Институт проблем управления сложными системами РАН, г. Самара
E-mail: minakov@magenta-technology.ru

Рассматривается задача интеллектуальной обработки информации, представленной в виде текста на естественном языке, с целью извлечения знаний, ориентированных на предметные интересы исследователя. Обсуждается архитектура системы, предназначенной для решения подобной задачи. Описываются основные блоки и возможности системы и рассматриваются перспективы применения данного подхода.

Постановка задачи. В современном обществе темпы роста количества информационных материалов все более увеличиваются. Большинство актуальных научно-технических проблем изучается независимо как в институтах и научных лабораториях, так и коммерческими предприятиями с целью их практического использования. Эксперту, желающему получить новую или систематизировать имеющуюся информацию об исследуемом объекте, потенциально доступно множество уже существующих источников информации, которая содержится в научных библиотеках, онлайн-конференциях, статьях и других источниках, включая Интернет.

К сожалению, на текущий момент инструментальных средств, позволяющих корректно получить с учетом знаний исследователя-эксперта информацию, ориентированную на эксперта и затрагивающую интересующую его предметную область, практически не существует.

Это связано с тем, что имеющиеся поисковые системы и системы документооборота работают с текстом документов (анализируя ключевые слова, морфологию, грамматику и т. д.), но не способны работать с его смыслом, так как не умеют анализировать знания, представленные в текстовой форме, что и является главной неразрешенной проблемой анализа.

Формализация подобного рода знаний – сложный процесс, связанный с их неоднородностью и противоречивостью, изменением и устареванием. Кроме того, зачастую даже человеку-эксперту (далее эксперту) трудно «извлечь» знания из собственного опыта и представить их как формализованное описание исследуемой предметной области (онтологию). Поэтому нужен инструмент, который мог бы помочь эксперту в построении подобного рода онтологий, создавая их первоначальный вариант автоматизированно, предлагая свои варианты и учитывая знания эксперта.

Но даже формализация знаний о предметной области не итог, а только первый шаг в подобном анализе. Необходимо иметь возможность представлять все документы – результаты исследований – в терминах такой онтологии (создавая семантические дескрипторы), иметь механизмы для сравнения, поиска и анализа таких дескрипторов, а также их классификации согласно содержащимся в них знаниям, иметь возможность интерактивного взаимодействия с экспертом при анализе, включая механизмы уточнения разработанной онтологии согласно результатам анализа.

В области теории и практики работы со знаниями накоплен значительный положительный опыт.

В то же время ситуация с созданием инструментальных средств для работы со знаниями значительно сложнее. Не только не существует единой инструментальной среды, обеспечивающей все шаги процесса приобретения знаний, но даже имеющиеся системы, ориентированные на решение отдельных подзадач, обладают целым рядом ограничений, существенно уменьшающих эффективность их практического использования.

Обзор программных решений, направленных на автоматизированное построение онтологий, можно найти в [1], средств обработки естественно-языковых текстов – в [2], методов семантической кластеризации – в [3] и методов пополнения онтологии – в [4].

В каждой группе программных систем можно выделить ряд принципиальных недостатков: необходимость существенной ручной предобработки данных экспертом; невозможность анализа всего набора текстов с точки зрения семантики предметной области; зависимость качества результатов от языка документов; отсутствие открытой модели предметной области, позволяющей в полной мере использовать знания эксперта и пополнять ее в процессе работы; ограниченность работы с семантическими сетями; непрозрачность и неинтерактивность алгоритмов; критичность к наличию «мусорной информации»; зависимость качества результатов от изначальной предпосылки-догадки о «правильной структуре»; нетерпимость к наличию неполной или противоречивой информации.

Поэтому задача интеграции знаний по-прежнему актуальна и разработка инструментальной системы интеграции профессиональных знаний, представленных на естественном языке, является важной.

Предлагаемый подход и архитектура системы. В данной работе представлена общая архитектура инструментальной системы, ориентированная на решение указанных выше задач и базирующаяся на предложенных в [5–7] технологиях понимания текстов на естественном языке и извлечения знаний на основе мультиагентного подхода [8]. Такая система позволяет анализировать наборы документов научно-технического содержания, представленные в виде текста на естественном языке, и получать предметно-ориентированную информацию согласно требованиям исследователя.

Каждый из модулей этой системы может использоваться автономно при решении конкретной практической задачи (например, при построении начальной онтологии логистики для определения типов объектов и их свойств перед процессом планирования или семантического мета-поиска документов в Интернете). Но все вместе они представляют собой завершенную среду анализа информации на естественном языке, реализуют все шаги, требуемые для такого анализа: конструирование начальных знаний, их анализ, систематизацию и замыкающее весь цикл пополнение новыми знаниями, полученными в результате анализа.

Предлагаемый подход к интеграции разнородных знаний, основанный на агентных взаимодействиях и заключающийся в совместном использовании агентных механизмов работы со знанием на естественном языке и мультиагентного кластерного анализа, позволил создать архитектуру работы со знанием для реализации предложенных методов автоматизированного конструирования онтологий, представления и обработки информации, анализа результатов и пополнения знаний, обеспечивая цикл познания, который необходим для эффективного и оперативного использования информации.

Разработанная среда состоит из нескольких программных комплексов: инструментария инженерии знаний, предназначенного для создания онтологий предметной области и логики принятия решений агентов, и программного инструментария, ориентированного на представление, анализ и обработку знаний, имеющихся в виде информации на естественном языке.

Архитектура системы приведена на рис. 1.

Инструментарий инженерии знаний включает в себя конструктор онтологий, автоматизированную систему построения онтологий, систему понимания текста на естественном языке, систему извлечения знаний, модуль пополнения онтологических знаний и ряд дополнительных модулей, в частности отладочную систему, интерфейсы работы с базами данных и внешними приложениями.

Таким образом, общая логика работы системы следующая: для получения новой информации об объекте исследования используется ряд документальных результатов, полученных другими экспертами (к таким результатам относится любая информация на естественном языке в электронном виде, включая документы, таблицы, электронную почту и т. д.).

На основе этой информации автоматизированно строится онтология предметной области, которая затем может быть уточнена и дополнена экспертом.

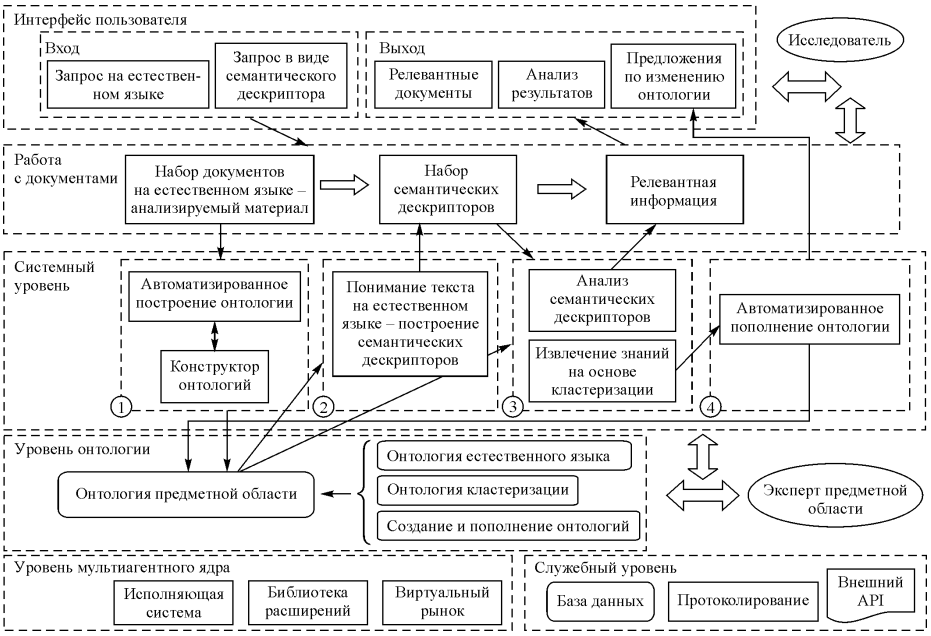

Рис. 1. Общая логическая архитектура системы

Для всех документов, содержащих результаты исследования, с применением технологии анализа текстов на естественном языке строятся семантические дескрипторы, которые позволяют представить смысл документов в терминах онтологии предметной области, т. е. в виде, удобном для семантического анализа.

Теперь модуль кластерного анализа может построить иерархические группы документов с учетом их семантической близости, а модуль анализа дескрипторов предоставит пользователю возможность с помощью интерфейса на естественном языке (также поддерживаются дескрипторы на основе онтологии) конструировать запросы и получать интересующую его информацию.

Далее работает модуль пополнения онтологии. Он использует определенные модулем анализа семантически близкие группы для анализа и последующего уточнения имеющейся онтологии найденными правилами, что в итоге дает возможность точнее описать предметную область и тем самым лучше представить семантику документов.

Таким образом, система одновременно работает в двух режимах: внешнем и внутреннем. Во внешнем режиме она ориентирована на конкретного исследователя и позволяет ему сконструировать онтологию, описывающую исследуемый объект, получить интересующие его знания, которые заключены в документах других экспертов, и проанализировать их с точки зрения семантической близости. Во внутреннем режиме система ориентирована на постоянное улучшение онтологии предметной области при использовании принципа обратной связи и пополнение онтологии найденными правилами и объектами.

Для создания специализированных программных компонент приложения предоставляется инструментарий, который состоит из расширяющегося набора библиотек программ и позволяет настроить инструментальную среду для обработки информации в исследуемой предметной области.

Рассматриваемые компоненты в целом составляют набор дополняющих друг друга средств, призванных упростить, ускорить и удешевить разработку подобного рода систем, а также обеспечить исследователю возможность эффективно и оперативно получать, обрабатывать и интегрировать знания об исследуемом объекте.

Опыт разработки и использования рассматриваемых систем показал высокую эффективность инструментальных средств, позволивших за короткое время создать более трех десятков различных приложений, предназначенных для решения реальных практических задач, в том числе в области логистики, конструирования онлайн-порталов, в поисковых системах, системах классификации документооборота и других применениях [9–12].

Модуль автоматизированного построения онтологии. В работе [7] предложен метод автоматизированного построения онтологии предметной области, реализуемый путем итеративного анализа строящейся онтологии с помощью методов понимания текстов на естественном языке с применением базовой онтологии и набора предметно-ориентированных текстов на основе алгоритмов мультиагентного взаимодействия и разрешения конфликтов между квантами знаний. Результатом этого является начальная онтология предметной области.

Задача автоматизированного построения онтологии включает в себя несколько этапов: выявление групп документов, относящихся к одной предметной области; выделение терминов предметной области на основе набора

документов; определение типа концепта онтологии для данного термина – деление терминов на объекты, отношения, свойства и атрибуты; анализ зависимостей между терминами (выявление свойств объектов, участников отношений и т. д.); выделение атрибутов и их значений; построение иерархической модели объектов; построение отношений в онтологии; этап «очистки» онтологии от незначимых терминов и проверка онтологии путем построения семантических дескрипторов документов и анализа противоречий.

Процесс построения онтологии не линейный, а итеративный – на каждом шаге возможен возврат на предыдущие и проверка на основе уже имеющейся онтологии интегральности созданной структуры. С использованием механизма понимания текстов на естественном языке документы разбираются по построенной онтологии, после чего проверяется корректность разбора на основе имеющихся правил и типовых шаблонов формирования онтологии.

На каждом этапе пользователь может быть вовлечен в процесс построения, давая дополнительные комментарии и правки и тем самым увеличивая достоверность найденных связей.

В работе [7] представлены алгоритмы, применяемые на каждом шаге построения онтологии, в том числе показано, как лингвистические шаблоны должны преобразовываться в онтологические конструкции, приведены механизмы распознавания значений атрибутов в тексте, а также описаны эвристические правила, позволяющие реконструировать зависимости между концептами в онтологии и отношения между объектами.

Особое внимание уделяется этапу проверки онтологии путем построения семантических дескрипторов документов и анализа противоречий, поскольку он является критическим для всей процедуры построения онтологии и представляет основное отличие предлагаемого подхода от известных методов, при этом являясь не независимым этапом, а постоянным процессом автоматического уточнения и верификации, запускаемым после каждого из этапов.

Проверка онтологии основана на предложенных алгоритмах автоматического понимания текста и является неким аналогом того, как на самом деле понимает текст незнакомой предметной области эксперт, а именно вначале строится предположение о возможном смысле того или иного концепта или группы концептов, затем на основе набора примеров, в которые этот концепт входит, проверяется правильность предположения как на основе синтаксических взаимосвязей, так и на основе семантических зависимостей

В рамках анализа синтаксиса проверяется, всегда ли концепт используется в той роли, которую ему приписали. Если нет, то является ли это случайностью (или ошибкой), другим термином, существующим наряду с первым, или же ошибкой в изначальной предпосылке. Таким образом, для каждого вхождения термина строится синтаксическое дерево отношений и определяется, корректно ли он «связался» с другими терминами. Так как на начальном этапе ошибка может быть и не в нем, то строится матрица связей между терминами и определяется степень корректности связи на основе правил грамматики языка. Далее все связи термина с другими терминами, у которых совокупная оценка качества их связей недостаточно высока, не учитываются. Тем самым отсекаются возможные ошибки неправильного понимания данного термина из-за ошибочной гипотезы о смысле другого термина. Для всех же «надежных» терминов оценивается качество связи при использовании данного термина в выбранном смысле. Если существует несколько возможных гипотез, то оцениваются все. Если одна из гипотез значительно лучше,

чем остальные, то термин считается надежным. В противном случае термин не считается надежным и его не рекомендуется использовать в качестве критерия при оценке надежности других терминов.

В рамках семантического анализа осуществляется подобный анализ, но уже на основе семантической онтологии предметной области. Оценивается: всегда ли концепт вступает в определенные семантические отношения; имеет ли целостность семантическая сеть, построенная с учетом данных предположений; нет ли противоречивых отношений с одним и тем же объектом; реконструируется ли все предложение в связный граф семантической сети или же в результате разбора получаются разрозненные концепты.

Для каждого термина проверяется: всегда ли он выступает в определенной ему роли (например, не выступает ли слово, которое определено как объект, в невозможной для него роли отношения); обладает ли стабильностью в наличии атрибутов и вступлении в отношения; не вступает ли в разных документах в противоречащие отношения или приобретает противоположные атрибуты одновременно.

Каждому концепту по итогам построения сцены для текста и оценки ее на связность и непротиворечивость назначается коэффициент в зависимости от того, насколько его связь с другими концептами повлияла на интегральные параметры. После этого анализируется общая точность термина на основе всего набора документов – насколько его участие повлияло на качество результатов. Если получаем значительный отрицательный коэффициент, то понимание смысла данного термина признается неудовлетворительным, так как в текущем своем значении он ухудшает понимание всего остального текста, и следует применить альтернативные его значения.

Модуль понимания текста на естественном языке. Предложен метод представления неструктурированной информации на естественном языке, который основан на применении механизмов мультиагентного взаимодействия квантов знаний, позволяющих реконструировать смысл предложения, и построенных онтологий для хранения межфразового контекста в виде семантических дескрипторов. Тем самым смысл текста представляется в виде семантических сетей и обеспечиваются механизмы сравнения семантики связных профессиональных текстов [5].

Суть предлагаемого подхода состоит в том, что каждому слову языка ставятся в соответствие агенты его смыслов, которые на основе собственных баз знаний (онтологий) конкурируют между собой и кооперируются, договариваясь о том, какой именно конкретный смысл имеет каждое слово в предложении и каков его общий смысл. В результате основной моделью процесса понимания смысла становится процесс самоорганизации смыслов слов при построении сцены контекста, что принципиально отличает предлагаемый подход от всех на сегодня известных.

Для реализации подхода разработана архитектура открытой мультиагентной системы [8], настраиваемой на различные применения и обеспечивающей возможность пополнения предметно-ориентированного словаря непосредственно в ходе ее работы. Система обеспечивает возможность морфологического и синтаксического анализа текста на естественном языке, понимания смысла и реализации практических действий на этой основе (прагматики).

Задачей морфологического анализа является разбиение строки на отдельные слова, поиск известных слов и их свойств в морфологической базе данных, а также поиск морфологических свойств неизвестных слов в имею-

щихся словарях. Был выбран подход анализа словоформ целиком, так как разбор слова и его параметров на основе набора эвристических правил словообразования индивидуален для каждого языка и требует очень серьезных временных затрат на настройку, в то время как определение словоформ можно автоматизировать.

Задачей этапа синтаксического анализа является выявление синтаксической структуры предложения, включающей в себя множество синтаксических ролей, соответствующих каждому слову, множество морфологических атрибутов слов, а также множество синтаксических зависимостей между отдельными словами. На этом этапе происходит поиск синтаксически совместимых пар слов и их объединение в словосочетания. Словосочетания, в свою очередь, также могут объединяться с другими словами (и словосочетаниями), в результате чего образуется древовидная структура, представляющая собой схему синтаксического разбора предложения. Онтология синтаксиса конструируется отдельно для каждого языка.

На смысловом этапе проверяются на непротиворечивость связи, выявленные на этапе синтаксического анализа. В результате взаимодействия слов образуется семантический дескриптор предложения, являющийся ориентированным графом, который содержит объекты, атрибуты и отношения, упомянутые пользователем во входной строке. Он образуется путем выбора из множества противоречащих понятий (если они есть) наиболее вероятных.

Новое предложение рассматривается с учетом имеющегося семантического дескриптора сцены, что позволяет уточнять (или изменять) понятия с учетом новой информации.

Пример построения семантического дескриптора для задачи [9] представлен на рис. 2.

Модуль кластеризации для извлечения знаний. Предложен метод кластерного анализа, реализованный на основе агентного взаимодействия, что обеспечивает механизмы динамической иерархии групп семантически схожих объектов как в пошаговом, так и в пакетном режиме, а также дает возможность работать с неструктурированными квантами информации, таким образом предоставляя механизм поиска, анализа и классификации знаний, содержащихся в неструктурированных текстах [6].

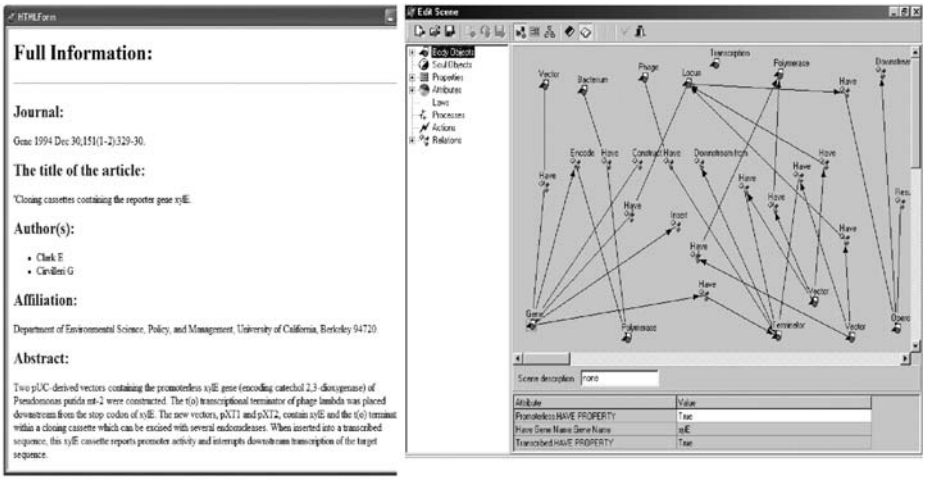

Рис. 2. Пример построения семантического дескриптора реферата статьи

В предлагаемом подходе в соответствие каждому элементу системы – каждой записи и кластеру – ставится программный агент, представляющий его интересы. Процесс работы системы состоит в переговорах, направленных на улучшение состояния элементов согласно критериям оценки качества. Вместо централизованной последовательной обработки осуществляется распределенная обработка, в которой каждая запись и каждый кластер самостоятельно и на основе некоторых заданных стратегий в узко ограниченном контексте принимают решения о вхождении в кластер или выходе из него, расширении или сужении кластера, или его удалении, тем самым представляя текущий локальный баланс интересов конкретных записей и кластеров. В итоге процесс кластеризации осуществляется путем самоорганизации агентов, формирующих иерархическую кластерную структуру.

Подобный взгляд на проблему кластеризации позволяет абстрагироваться от реальной структуры данных и работать с понятиями более высокого уровня, что дает возможность предложить универсальный алгоритм, в равной мере ориентированный как на структурированные данные (таблицы, базы данных), так и неструктурированные (текст на естественном языке). Также к дополнительным плюсам относится естественная параллельность, обеспечиваемая принципами построения архитектуры мультиагентных систем, возможность гибко изменять стратегии и отслеживать логику принятия решений, а также поддержка работы в режиме реального времени за счет динамической перестройки кластерной структуры с учетом вновь приходящих записей.

Общая схема работы алгоритма следующая: в момент появления каждой новой записи ей назначается агент и дается некоторая начальная сумма денег (энергия), используя которую запись пытается вступить в различные кластеры. На основе критерия ценности кластеров и расстояния до них запись определяет те кластеры и записи, с которыми она хотела бы объединиться. При этом целью записи является максимизация суммарной ценности кластеров, в которых она состоит (деньги могут браться за вход в кластер и/или покрытие расстояния до него).

В процессе переговоров определяются те кластеры и записи, которые готовы принять новую запись. Переговоры необходимы, так как предложение записи может быть невыгодно кластеру.

После вхождения записи в ряд кластеров и/или изменения их иерархической структуры аналогичные переговоры осуществляют те записи и кластеры, чья ценность изменилась (например, увеличилось число записей, изменилось описание полей, выросла энергия).

Целью кластера является повышение собственной ценности, а также повышение его суммарной ценности за счет вступления в другие кластеры более высокого уровня.

Формула ценности выбирается в зависимости от конкретной задачи и является функцией от таких параметров, как число записей в кластере (и общее число записей в системе), объем кластера, число атрибутов, по которым образован кластер (и общее число атрибутов), пределы изменения атрибутов, распределение значений по атрибутам, степени влияния атрибутов (т. е. кластеры по одним атрибутам более ценны, чем по другим). В простейшем случае ценностью кластера можно считать его локальную плотность. Для случая кластеризации семантических сетей ценность определяется наличием конкретных концептов онтологии, количеством концептов, их отношениями и степенью онтологической близости между ними.

Описанием записи является либо строка (запись) в базе данных (список атрибутов с конкретными значениями), либо семантическая сеть (в случае неструктурированных данных). Описанием кластера будет соответственно либо список атрибутов с допустимыми диапазонами – границами кластера, либо семантическая сеть, представляющая общую часть (онтологическое пересечение) записей, входящих в кластер.

Записи, которым кластер предлагает войти в него, потенциально могут уже состоять в некоторых кластерах, но если предлагаемый вариант лучше, то запись может выйти из старого кластера и вступить в новый, перераспределяя имеющуюся у нее энергию. В результате ценность старого кластера уменьшается (потенциально другие записи тоже могут покинуть его) и в итоге кластер может разрушиться.

Процесс поиска вариантов заканчивается, когда все элементы системы нашли устраивающие их варианты, либо за круг переговоров никто из записей и кластеров не сумел договориться и изменить свое состояние.

Данный процесс ограничивается вследствие уменьшения уровня энергии вовлеченных кластеров и записей, поэтому он затухающий. Структура кластеров после каждого воздействия стабилизируется, суммарно улучшая свою ценность, тем самым отражая динамически изменяющуюся картину и представляя все более значимые зависимости для эксперта.

В [6] описываются типовые стратегии записи и кластера, поддерживаемые типы полей, возможные способы представления структуры кластеров, методы вычисления расстояний между записями и кластерами, формулы ценности для кластера и записи, принципы точной и интервальной кластеризации, преобразование и нормирование атрибутов, параметры микроэкономики, в том числе назначение начального количества денег, механизмы поиска вариантов вхождения в кластер, распределение денег между кластерами, выход из кластера и налоги.

Модуль автоматизированного пополнения онтологии. Предложен метод автоматизированного пополнения онтологии новыми знаниями на основе анализа семантических групп, найденных на этапе кластеризации, и применения ряда эвристических правил, позволяющих уточнить и пополнить онтологию предметной области, улучшая таким образом качество представления, поиска и анализа документов [13].

Модуль автоматизированного пополнения онтологии дает возможность на основе найденных групп семантически близких дескрипторов «выращивать» новые связи между существующими в онтологии квантами знаний.

Пополнение и уточнение онтологии основано на гипотезе взаимодействия: если концепты онтологии всегда встречаются вместе в определенной ситуации, значит, они семантически связаны между собой, причем характер связи определяется ситуацией. С помощью методов модуля можно проанализировать получившуюся структуру и дескрипторы кластеров и выделить не обнаруженные ранее зависимости между концептами онтологии (например, два объекта в онтологии должны быть связаны неизвестным отношением, так как всегда встречаются вместе, или два атрибута на самом деле являются дублем одного и того же свойства). Данный процесс может проходить как автономно, так и в интерактивном диалоге с пользователем.

После того как документы получили семантические дескрипторы и кластеризованы по семантической близости, происходит процесс кластеризации созданных ранее кластеров. Теперь анализируются те зависимости, по которым были объединены документы в различных группах. Подобный процесс

позволяет подняться над уровнем документов и исследовать уже саму предметную область, анализируя концепты, встречающиеся в различных семантически близких группах, а также установить возможные взаимосвязи между ними. Естественно, что для корректных гипотез требуется большая выборка документов исследуемой предметной области (простое эвристическое правило – анализируемых документов должно быть на порядок больше, чем концептов в исследуемой онтологии).

В результате по итогам анализа семантики кластеров для каждой группы (кластера кластеров) определяется ряд возможных пополнений (или изменений) в онтологии. При этом для каждого из вариантов изменения считается степень его корректности путем временного изменения онтологии и анализа числа корректных/некорректных использований измененной части онтологии на имеющемся наборе документов. Все варианты и их степень корректности предлагаются пользователю, который может в интерактивном режиме изменить и уточнить предложенные гипотезы для окончательного утверждения и пополнения онтологии.

Выявлены следующие типовые комбинации концептов онтологии: два несвязанных объекта; два объекта, связанные определенным отношением; два объекта, всегда связанные двумя конкретными отношениями; объект плюс другой объект, связанный определенным отношением с различными третьими объектами; объект связан отношениями одного и того же типа с двумя объектами разных типов; объект плюс атрибут, встречаемый у других различных объектов; объект плюс атрибут, всегда наличествующий у другого объекта; объект плюс отношение, не связанные ни с каким объектом; два атрибута, встречающиеся у одного и того же объекта; один атрибут, встречающийся одновременно у нескольких разных объектов (в случае устойчивой комбинации).

Для каждой комбинации в [13] описаны возможные варианты пополнения (или изменения) онтологии, а также способы проверки корректности гипотез.

Предлагаемый подход к интеграции профессиональных знаний позволяет добиться следующих основных преимуществ перед существующими методами:

1. Алгоритмы не требуют предобработки и фильтрации данных экспертом предметной области, а также участия человека в процессе работы, но могут использовать интерактивное взаимодействие с экспертом для повышения качества результатов.

2. Процесс принятия решений (от построения онтологии до создания семантических дескрипторов и формирования кластеров) полностью прозрачен для пользователя. Обоснования всех принимаемых решений, логика и оценки могут быть прослежены.

3. Поддерживается открытая модель предметной области, позволяющая в полной мере использовать знания эксперта, давая ему мощный инструмент настройки и пополнения онтологии знаниями о предметной области в процессе работы.

4. Имеется возможность представления смыслового контекста связного текста за счет использования механизмов представления и обработки знаний на основе онтологий предметных областей с поддержкой уточнений, разрешением противоречий и т. д.

5. Имеется возможность кластерного анализа на основе онтологий, что позволяет кластеризовать сложные информационные объекты (образы, тексты) с учетом их семантики.

6. Поддерживается описание кластера в терминах онтологии, что дает возможность удобного анализа результатов, описание кластера правилом вида «если–то», создание значимых кластеров в любом подпространстве исследуемого пространства решений.

7. Поддерживается возможность работы с множеством документов из нескольких слабосвязанных предметных областей за счет предварительного этапа автоматической предобработки алгоритмом кластеризации.

Пример использования инструментальной среды для решения проблемы семантико-ориентированного поиска в информационно-поисковой системе MEDLINE. База данных MEDLINE хранит рефераты научных статей из области биологии, химии и медицины. Имеющийся механизм поиска по ключевым словам не удовлетворял качеством результатов. Требовался новый механизм семантического поиска для описания рефератов и запросов на их поиск в виде семантических сетей (по аналогии с технологией Semantic Web), который позволит повысить качество поиска информации.

Решение данной задачи осуществлялось в несколько этапов. На первом этапе на основе знаний экспертов была построена онтология предметной области, которая должна описывать возможные семантические понятия, встречающиеся в интересующих статьях выбранной предметной области (например, молекулярной биологии). Ряд статей был проанализирован с помощью построенной онтологии, в результате чего она пополнилась недостающими понятиями и отношениями.

Вторым этапом стал процесс построения семантического дескриптора для каждой статьи, полученной на основе предварительного поиска по ключевым словам. Этот процесс был автоматизирован с помощью технологии понимания текста на естественном языке (см. рис. 2).

На третьем этапе были выполнены анализ и сравнение полученных семантических дескрипторов с дескриптором запроса и выявление релевантных статей. Осуществлен анализ текстов статей и сравнение семантики статьи с запросом. Были проанализированы причины ошибок и в рамках четвертого этапа осуществлено пополнение онтологии с тем, чтобы точнее реконструировать семантические дескрипторы документов за счет новых концептов онтологии и уточнения отношений между ними.

Весь процесс был повторен несколько раз, пока не удалось добиться существенного повышения качества возвращаемых статей на любой запрос в рамках исследуемой предметной области. В частности, удалось повысить семантическую релевантность результатов поиска на 80 % и скорость решения практических экспертных задач примерно в 6 раз (например, задача выбора статей в области молекулярной биологии, описывающих определенные типы экспериментов с требуемыми параметрами результата, заняла 8 человеко-месяцев вместо изначально планировавшихся по оценкам экспертов 4 человеко-лет) [9].

Заключение. Предложенный подход и разработанная инструментальная система для решения задач извлечения знаний и понимания текста на естественном языке предоставляют исследователю удобные и развитые механизмы для анализа разнородной информации в виде электронных информационных ресурсов.

Инструмент получения информации и знаний, ориентированных на конкретного человека-эксперта, учитывает как общие знания о предметной области, так и его личную модель понимания и является неоценимым подспорьем в любых научно-технических и коммерческих исследованиях.

СПИСОК ЛИТЕРАТУРЫ

1. **Gomez-Perez A., Manzano-Macho D.** A Survey of Ontology Learning Methods and Techniques. Deliverable 1.5, OntoWeb Project, 2003.
2. **Шаров С. А.** Средства компьютерного представления лингвистической информации. Обзор // http://www.kcn.ru/tat_en/science/ittc/vol000/002/
3. **Steinbach M., Karypis G., Kumar V.** A comparison of document clustering techniques // Proc. of KDD Workshop on Text Mining. 2000.
4. **Mitchell T.** Machine Learning. McGraw Hill, 1997.
5. **Андреев В. В., Ивкушкин К. В., Карягин Д. В. и др.** Разработка мультиагентной системы понимания текста // Тр. Третьей междунар. конф. по проблемам управления и моделирования сложных систем. Самара: СНЦ РАН, 2001. С. 489.
6. **Андреев В. В., Волхонцев Д. В., Ивкушкин К. В. и др.** Мультиагентная система извлечения знаний // Там же. С. 206.
7. **Минаков И. А.** Разработка автоматизированной системы построения онтологии предметной области на основе анализа текстов на естественном языке // Вестн. Самар. гос. техн. ун-та. Сер. Технические науки. 2004. Вып. 20. С. 44.
8. **Андреев В., Батищев С., Виттих В. и др.** Методы и средства создания открытых мультиагентных систем для поддержки процессов принятия решений // Изв. РАН. Сер. Теория и системы управления. 2003. № 1.
9. **Андреев В., Гельфанд М., Ивкушкин К. и др.** Мультиагентная система для интеллектуального поиска рефератов статей по молекулярной биологии // Тр. Четвертой междунар. конф. по проблемам управления и моделирования сложных систем. Самара: СНЦ РАН, 2002. С. 338.
10. **Андреев В., Минаков И., Лахин О. и др.** Развитие элементов самоорганизации и эволюции в мультиагентном портале социокультурных ресурсов Самарской области // Тр. Шестой междунар. конф. по проблемам управления и моделирования сложных систем. Самара: СНЦ РАН, 2004. С. 277.
11. **Алексеев А., Вольман С., Минаков И. и др.** Создание мультиагентной системы автоматической обработки, преобразования и коррекции логистических сообщений стандартных форматов обмена бизнес-данными // Там же. С. 270.
12. **Андреев В., Вольман С., Ивкушкин К. и др.** Разработка мультиагентной системы интеллектуальной обработки и классификации документов // Тр. Пятой междунар. конф. по проблемам управления и моделирования сложных систем. Самара: СНЦ РАН, 2003. С. 317.
13. **Минаков И. А.** Автоматизированное пополнение онтологии на основе знаний, извлеченных в процессе кластеризации // Вестн. Самар. гос. техн. ун-та. Сер. Технические науки. 2005. Вып. 33. С. 321.

Поступила в редакцию 8 декабря 2005 г.