Г.В. Рыбина, А.О. Дейнеко

Распределенное приобретение знаний для автоматизированного построения интегрированных экспертных систем[1]

Аннотация. Обсуждается проблема распределенного приобретения знаний для построения полных и непротиворечивых баз знаний в интегрированных экспертных системах за счет совместного использования источников знаний различной типологии (эксперты, проблемно-ориентированные тексты, электронные носители в виде баз данных). Основное внимание уделяется моделям, методам и алгоритмам распределенного приобретения знаний из баз данных как дополнительного источника знаний. Приводится описание архитектуры и базовых функциональных возможностей средств распределенного приобретения знаний, функционирующих в составе инструментального комплекса АТ-ТЕХНОЛОГИЯ.

Ключевые слова: интегрированные экспертные системы, распределенное приобретение знаний, база знаний, база данных, продукционные правила.

Введение

Проблема приобретения знаний всегда находилась в центре внимания разработчиков современных интеллектуальных систем, в частности, традиционных экспертных систем (ЭС) и более сложных – интегрированных экспертных систем (ИЭС), обладающих масштабируемой архитектурой и расширяемой функциональностью [1]. Этому важнейшему направлению искусственного интеллекта посвящено значительное число исследований и разработок, результаты которых широко представлены в фундаментальных зарубежных, например [2], и отечественных работах [1, 3-5].

Тем не менее, вопросы практического использования традиционных методов приобретения знаний и создания технологии автоматизированного приобретения знаний попрежнему являются актуальной проблемой, что связано как с острым дефицитом экспертов, так и с нехваткой специальных компьютерных систем, имитирующих искусство эксперта/экспертов. Исследования, проведенные в когнитивной психологии, показывают, что путь от новичка до эксперта в любой области профессиональной деятельности составляет в среднем около 10 лет, т. е. требуется достаточно длительная профессиональная практика, чтобы специалист выработал подходы (правила) принятия успешных решений [6].

Наиболее остро проблема приобретения знаний возникает при решении сложных практических задач, особенно в таких областях, как медицина, энергетика, космос, экология и др., где не всегда достаточно мнения одного эксперта, поэтому для построения максимально полных и непротиворечивых моделей проблемных областей (ПрО) и снижения рисков ошибок эксперта необходимо привлекать нескольких экспертов или группу экспертов, что существенно удорожает стоимостные и временные параметры разработки ИЭС [1]. Соответственно возрастает актуальность и роль степени автоматизации труда экспертов и разработки специальных программных средств, различных «оболочек приобретения» и т.д., направленных на компьютерную поддержку процессов получе-

[1] Работа выполнена при поддержке РФФИ (проект №09-01-00638)

ния знаний от эксперта или групп экспертов, являющихся основным источником знаний (источник знаний 1-ого типа [5]).

Однако, в настоящее время существует достаточно небольшое число исследований в области группового извлечения знаний от экспертов, наиболее известными из которых являются работы [6, 7], которые пока носят только теоретико-методологический характер, а из зарубежных – это проект [8], описывающий возможности графического представления распределенных знаний, а также работы французской группы ACACIA по созданию инструментального средства KATEMES (Knowledge Acquisition Toot for Explainable, Multi-Expert Systems) [9], предназначенного для частичной автоматизации работы инженера по знаниям на этапе группового извлечения знаний.

С другой стороны, типология источников знаний уже не ограничивается только экспертами. Значительные объемы экспертных знаний накоплены в текстах на естественных языках (источник знаний 2-ого типа). В последние годы возник источник знаний 3-его типа, т.е. знания из современных информационных систем, представляющих собой сложные организационно-технические системы с такими компонентами управления как сетевые устройства, серверы, приложения, БД (СУБД) и т.д.

Проблема получения (выявления) знаний из источников 2-ого типа связана с бурно прогрессирующей технологией Text Mining [10], а проблеме автоматизированного извлечения знаний из БД в искусственном интеллекте посвящены такие новые направления как Data Mining и Knowledge Discovery in Databases (KDD) [11]. Успехи технологии Text Mining связаны с различными аспектами применения текстологических методов получения знаний из естественно-языковых текстов (ЕЯ-текстов), которые получили наибольшее развитие в трёх типах современных веб-ориентированных ЕЯ-систем – поиска информации (Information Retrival), извлечения информации (Information Extraction) и понимания ЕЯ-текста (Text / Message Understanding) [12].

С применением различных алгоритмов Data Mining тесно связаны такие ПрО, как: научные исследования (медицина, биология, биоинформатика и др.); решение задач бизнеса (банковское дело, финансы, страхование, CRM и др.); задачи государственного уровня (борьба с терроризмом, поиск разыскиваемых лиц и т.д.); решение задач анализа веб-ресурсов, где основными направлениями являются Web Content Mining (интеллектуальные поисковые агенты, классификация и фильтрация информации) и Web Usage Mining (подразумевает обнаружение закономерностей в действиях пользователя веб-узла или их группы) и др.

Каждая их этих технологий возникла и развивалась независимо друг от друга, и сегодня подобная *автономность* и *распределенность* не позволяет осуществлять эффективный мониторинг всех информационных ресурсов (базы знаний, базы данных, а в последние годы и онтологии), которыми обладают интеллектуальные системы, в частности ИЭС. В настоящее время, практически, отсутствуют исследования в области создания инструментальных средств и технологий распределенного приобретения знаний из различных источников различной типологии.

Опыт практического использования целого ряда прикладных ИЭС, разработанных на основе задачно-ориентированной методологии (ЗОМ) и инструментального комплекса АТ-ТЕХНОЛОГИЯ [1] (в том числе, для экспресс-диагностики крови, диагностики сложных технических систем, проектирования уникальных объектов машиностроения, комплексных экологических задач и др.), показал необходимость *мониторинга*, т.е. проведения регулярных проверок и подтверждений, накапливаемых и формализуемых знаний в соответствующих базах знаний (БЗ).

Кроме выявления ошибок (дефектов), дублирования, противоречивости и неполноты информации в БЗ функционирующих систем, эти же вопросы имеют важное значение при моделировании ПрО и проектировании собственно БЗ и БД (контроль ограничений целостности, согласованности, соглашений между использованием терминов ПрО и т.д.). Например, как показано в [1], чтобы преодолеть проблему *неполноты* разрабатываемой БЗ (т.е. эксперт не знает и/или забыл отметить какой-либо факт, необходимый для решения задачи) можно поступать следующим образом: приглашать конкретного эксперта n-ое количество раз; приглашать других экспертов или группу экспертов; использовать независимый электронный источник знаний в виде БД. Первые два способа могут привести к срыву всего процесса моделирования ПрО как из-за суще-

ственного удорожания стоимости труда эксперта/экспертов, так и в следствии так называемых «шумовых» личностных особенностей экспертов (недопонимание, умолчание, конформизм, когнитивная защита, собственные интересы эксперта, отсутствие семантической унификации используемых терминов ПрО и др. [5]). В [3] также особо отмечается наличие таких факторов, как «когнитивная защита личности», «дискретность», неполнота человеческого знания и др.

Наиболее нейтральными и независимыми источниками знаний являются БД. Анализ экспериментальных данных, полученных при создании БЗ целого ряда прикладных ИЭС, показал, что *локальное* использование БД в качестве дополнительного источника знаний способно пополнить объем разрабатываемых БЗ на 10-20%, в зависимости от специфики ПрО [1].

Таким образом, возникает необходимость создания новой автоматизированной технологии приобретения знаний, распределенных по различным источникам. Можно сформулировать следующий концептуальный базис, положенный в основу данной работы:

1. вводится понятие «распределённого приобретения» знаний применительно к интеграции информации, полученной из источников знаний различной типологии;

2. источники знаний 1-го и 2-го типа в контексте комбинированного метода приобретения знаний (КМПЗ) [1], реализованного в рамках ЗОМ, рассматриваются как *совмещенные*, поскольку в КМПЗ существует совокупность хорошо апробированных технологических процедур, позволяющих дополнять информацию, полученную от эксперта/экспертов, за счет информации, выявленной из проблемно-ориентированных ЕЯ-текстов (в данном случае – это обработка протоколов интервьюирования экспертов, сбор лексики инженера по знаниям / системного аналитика, анализ сигнальных лексем во входных ЕЯ-текстах и др.);

3. в центре внимания *распределённого* приобретения знаний – проблема интеграции с информацией, полученной из БД как источника знаний 3-го типа с целью автоматизированного построения максимально полных и непротиворечивых моделей ПрО;

4. поскольку не существует универсальных методов, позволяющих решать проблему *неполноты* БЗ, то разработка и применение технологии приобретении знаний из БД как дополнительного источника знаний является достаточно новым приложением концепций Data Mining и KDD для решения этой проблемы.

В данной работе обсуждаются модели, методы и алгоритмы распределённого приобретения знаний на основе KDD и Data Mining и описывается типовая проектная процедура технологии применения KDD и Data Mining на различных стадиях жизненного цикла, связанного с автоматизированным построением БЗ прототипов ИЭС.

1. Общая характеристика комбинированного метода приобретения знаний

В соответствии с концептуальными основами ЗОМ построения ИЭС неотъемлемой частью данной методологии является ЗОМ приобретения знаний, представляющая собой совокупность КМПЗ и технологии его использования на различных стадиях жизненного цикла построения ИЭС и веб-ИЭС [1]. В рамках базового КМПЗ и средств его реализации рассматривается так называемый *локальный вариант* приобретения знаний.

Однако при переходе к веб-версии комплекса АТ-ТЕХНОЛОГИЯ стал возможен другой вариант автоматизированного приобретения знаний на основе КМПЗ – *распределенный*, обеспечивающий в рамках клиент-серверной архитектуры, с одной стороны, интеграцию всех рассмотренных выше типов источников знаний, с другой стороны – учет их географической распределённости, а также возможность работы с группами удаленных источников знаний.

В целом обобщенную модель КМПЗ [1] с учетом особенностей распределенного приобретения знаний можно представить в виде:

$$M_{км} = <N^{\sim}, S^{\sim}, F^{\sim}, K, Z>,$$

где $N^{\sim} = \{N^{\sim}_{лок\ n}\}$, $n=1,\ldots,m_n$ – множество неструктурированных описаний ПрО;

$N^{\sim}_{лок\ n} = <IN, TN, SN, CN>$, где IN – порядковый номер описания; TN – тип описания, SN – источник, откуда получено описание, CN – собственно само описание;

$S^{\sim} = \{S^{\sim}_{m}\}$, $m=1,\ldots,m_m$ – множество структурированных описаний ПрО;

$F^{\sim}$ – множество процедур отображения $N^{\sim}$ в $S^{\sim}$;

K – процедуры конвертации сформированного поля знаний (ПЗ) в форматы языков представления знаний (ЯПЗ) различных инструментальных средств для построения ЭС (зарегистрированных в комплексе АТ-ТЕХНОЛОГИЯ);

Z – фрагменты БЗ в форматах ЯПЗ других инструментальных средств построения ЭС.

Следовательно, в ходе сеанса интервьюирования эксперта осуществляется *структурирование* полученной информации в виде ПЗ, выполняющего важную функцию в процессе структурирования полученной от эксперта информации о ПрО, обеспечивая единое внутреннее представление и унификацию основных понятий и отношений ПрО, выявленных из различных источников знаний как первый шаг к формализации на конкретном ЯПЗ.

Соответственно, с учетом особенностей распределенного приобретения знаний обобщенную модель ПЗ можно представить в виде:

$$S^{\sim}_m = <IS_m, TS_m, SS_m, O_m, R_m>,$$

где IS_m – порядковый номер структурированного описания ПрО;

TS_m – тип структурированного описания ПрО;

SS_m – источник, откуда получено описание;

$O_m = \{O_{mj}\}$, j=1,…,n – множество объектов;

$R_m = \{R_{mk}\}$, k=1,…,p – множество правил.

Таким образом, при переходе от локального варианта приобретения знаний к распределенному варианту множество базовых процедур КМПЗ пополняется следующими процедурами:

– получение описаний из распределенных источников;

– сопоставление приобретенных знаний разного типа;

– уточнение описаний с выявленными несоответствиями;

– групповое извлечение знаний.

2. Особенности применения KDD для распределенного приобретения знаний

Как уже отмечалось выше, для приобретения знаний из БД в рамках КМПЗ используются технологии KDD и Data Mining, применение которых в качестве *дополнительного источника знаний* для преодоления неполноты БЗ обеспечивает интеллектуальный анализ больших объемов информации и выявление в них скрытых закономерностей в ИЭС, разрабатываемых на основе ЗОМ.

Следует отметить, что в ЗОМ эти термины трактуются следующим образом: под KDD подразумевается *весь процесс* извлечения знаний, начиная от соединения с БД, заканчивая представлением полученных результатов, а Data Mining являешься лишь некоторым *этапом* общего процесса KDD.

С точки зрения процессов приобретения знаний концепция Data Mining реализована в КМПЗ тремя следующими способами [1]: генерация начального ПЗ из БД с последующей модификацией его экспертом; верификация ПЗ, полученного в процессе интервьюирования эксперта, а также его частичная модификация, связанная с нахождением коэффициентов уверенности для уже выявленных знаний; слияние ПЗ, полученных в результате применения двух методологий.

Одной из особенностей применения KDD и Data Mining в рамках КМПЗ является необходимость организации доступа к конкретной БД, содержащей информацию по анализируемой ПрО, а также ее предобработки. Поэтому КМПЗ включает в себя множество специальных процедур для работы с БД, таких как:

– генерация SQL-запроса к СУБД;

– извлечение данных из БД в соответствии с запросом, сформированным процедурой извлечения данных из БД;

– фильтрация некоторого подмножества данных, которое в дальнейшем будет использоваться для построения набора правил (процедура фильтрации подмножества данных);

– преобразование данных для конвертации в формат, который может напрямую использоваться алгоритмами приобретения знаний (процедура преобразования данных).

Ниже приводится описание процедур, предназначенных для подготовки выборки данных для последующего анализа.

На основе процедуры генерации SQL-запроса формируется выборка для дальнейшего применения алгоритмов Data Mining. Инженер по знаниям выбирает атрибуты из БД, включаемые в выборку, на основании которой система генерирует SQL-запрос. С учетом специфики используемых в КМПЗ алгоритмов Data

Mining с помощью инженера по знаниям осуществляется процедура выделения зависимых и независимых атрибутов (столбцов) в анализируемой выборке, затем происходит обработка неизвестных значений атрибутов.

Следует отметить, что в локальном варианте КМПЗ использовались два базовых алгоритма построения деревьев решений ID3 [13] и C4.5 [14], позволяющих путем анализа построенных деревьев решений строить наборы продукционных правил. Однако, при переходе к распределенному варианту КМПЗ предпочтение было отдано концепции алгоритма CART [15], позволяющего строить бинарные деревья решений, что более удобно при визуализации и возможной постобработке выведенных правил, направленной на уменьшение общего количества выведенных правил.

Процедура преобразования данных осуществляет конвертацию в формат, который может напрямую использоваться алгоритмами приобретения знаний. После того, как выборка для анализа готова, применяется непосредственно процедура приобретения знаний из БД, обеспечивающая определение зависимостей в виде продукционных правил и использующая тот или иной алгоритм.

Заключительными являются три следующих процедуры: оценка точности полученной модели с использованием тестовых данных; определение алгоритма и его параметров, обеспечивающих наилучший результат в процессе приобретения знаний, и конвертация полученных правил в необходимый формат.

При переходе к распределенному варианту приобретения знаний особое внимание уделяется синхронизации процессов получения знаний из различных источников, что обеспечивается с помощью специальной типовой проектной процедуры (ТПП) «Приобретение знаний из БД», предусмотренной в ЗОМ и в технологии построения прототипов ИЭС [1]. Применяемая ТПП использует технологическую БЗ интеллектуального планировщика комплекса АТ-ТЕХНОЛОГИЯ и специальные программные средства для интеграции источников знаний, на основе которых осуществляется объединение фрагментов ПЗ, получаемых из разных источников.

Сценарий выполнения ТПП «Приобретение знаний из БД» включает в себя следующие этапы:

- получение фрагментов ПЗ в виде наборов продукционных правил за счет использования КМПЗ (интервьюирование экспертов, приобретение знаний из БД) и проведение последующей верификации полученных фрагментов ПЗ;
- программное объединение наборов правил за счет реализации алгоритма сравнения нескольких фрагментов ПЗ, основанного на расчете коэффициента меры близости [16] для каждой пары участвующих в сравнении правил;
- верификация единого ПЗ.

Отметим, что объединение наборов правил является одной из наиболее трудоемких задач. Этой процедуре предшествует автоматизированное сравнение наборов правил, полученных из разных источников знаний [17]. В качестве анализируемой структуры для эффективного и быстрого сравнения наборов правил в ЗОМ используются расширенные таблицы решений (РТР) [18], представляющие собой набор строк и столбцов, где каждая ячейка строки РТР хранит данные о вхождении и параметрах вхождения утверждения, характеризующегося заголовком строки, в конкретное правило.

Ниже приводится описание разработанных алгоритмов приобретения знаний из БД на основе построения бинарных деревьев решений и сравнения продукционных правил.

3. Алгоритм приобретения знаний из БД на основе построения бинарных деревьев решений

Алгоритм приобретения знаний из БД разработан на основе алгоритма CART [15] и представляет собой алгоритм построения бинарных деревьев решений. Для построения дерева решений на каждом узле дерева необходимо осуществлять разбиение множества атрибутов, ассоциированного с узлом, на два подмножества. В идеале выбранный атрибут должен разбить множество таким образом, чтобы два полученных подмножества состояли из объектов, принадлежащих к одному классу. Однако на практике разбиения такого качества получаются редко, поэтому правило можно сформулировать следующим образом: при разбиении множества атрибутов, ассоциированного с текущим узлом дерева, выбранный атрибут должен разбивать множество на подмножества таким образом, чтобы количество объектов из других классов

было минимальным. В алгоритме приобретения знаний из БД для оценки качества разбиения используется статистический критерий, основанный на индексе Gini [15].

На первом шаге работы алгоритма загружается контрольная выборка – предварительно подготовленная плоская таблица данных. Все данные записываются в нее, и на основе значений индекса Gini формируется корневой узел дерева, а сама матрица разбивается на две части: левую (не выполняется правило разбиения) и правую (выполняется правило разбиения). В случае если примеры на некотором конечном узле (листе) дерева принадлежат одному классу, то данный лист помечается именем класса. Если не все листья дерева помечены именами классов, то проверяется, все ли примеры в узле принадлежат одному классу, если нет, то проводится дальнейшее разбиение. Если все листы помечены именами классов, то дерево считается построенным. Далее, путем обхода дерева генерируется набор готовых продукционных правил, причем каждый путь от корня до листа дерева даёт одно правило, а условиями правила являются проверки из узлов, принадлежащих пути.

Важно отметить, что алгоритмы приобретения знаний зачастую генерируют множество простых продукционных правил, которые могут быть преобразованы в одно или несколько более сложных, но в то же время более наглядных и удобных для использования правил. Для решения этой проблемы в рассматриваемом алгоритме применяется постобработка сгенерированных правил, направленная на сокращение их количества. Способы упрощения правил можно разделить на четыре общих группы:

Способ 1. Некоторый набор выведенных правил имеет одинаковое заключение. В этом случае посылка нового правила является дизъюнкцией посылок рассматриваемого набора правил. Например, пара правил вида: x_1=a & x_2=1 => x_3=1 и x_1=a & x_2=2 => x_3=1 могут быть преобразованы в правило вида x_1=a & (x_2=1| x_2=2)=> x_3=1.

Способ 2. Некоторый набор выведенных правил имеет одинаковое заключение. В этом случае посылка нового правила является конъюнкцией посылок рассматриваемого набора правил, причем результирующее правило является более сильным, чем исходные. Например, пара правил вида: x_1 & x_2 => x_3 и x_1 & x_4 => x_3 могут быть преобразованы в правило вида x_1 & x_2 & x_4 => x_3. Следует учитывать, что посылки первоначального набора правил могут быть взаимоисключающими, в таком случае выполнение преобразования не возможно.

Способ 3. Некоторый набор выведенных правил имеет одинаковые посылки. В этом случае заключение нового правила является дизъюнкцией заключений рассматриваемого набора правил. Например, пара правил вида: x_1 & x_2 => x_3 и x_1 & x_2 => x_4 могут быть преобразованы в правило вида (x_1 & x_2) => x_3| x_4. Данный способ носит скорее теоретический характер, т.к. при решении задач классификации необходимо находить условия, четко относящие пример к тому или иному классу. Хотя иногда такое объединение может быть оправданно, но чаще нужно оставлять лишь одно из выведенных правил.

Способ 4. Некоторый набор выведенных правил имеет одинаковое заключение и в посылках правил рассматриваются числовые атрибуты. В этом случае возможно объединение условий, налагаемых на числовые атрибуты. Например, тройка правил вида: x_1=1 => x_4 и x_1=2 => x_4 и x_1=3 => x_4 может быть преобразована в правило вида ($x_1 \geq 1$ && $x_1 \leq 3$)=> x_4. Данный способ требует дополнительных знаний, в частности, о дискретных значениях атрибутов.

Перечисленные способы в некоторых случаях позволяют существенно уменьшить объём сгенерированных изначально правил. Полученные в результате постобработки правила должны оцениваться на тестовых данных и в окончательном наборе правил, предоставляемом для верификации эксперту, нужно оставлять только те правила, качество которых выше некоторого заданного порогового значения.

4. Алгоритм объединения наборов правил

Как уже отмечалось выше, в качестве анализируемой структуры для эффективного и быстрого объединения наборов правил в ЗОМ используются расширенные таблицы решений (РТР), представляющие собой набор строк и столбцов, где каждая ячейка строки РТР хранит данные о вхождении и параметрах вхождения *утверждения*, характеризующегося заголовком строки, в конкретное *правило*. Каждая ячейка РТР разбита на 2 части: одна – для IF-частей

правил, а другая – для THEN-частей правил. Обе части имеют одну и ту же структуру, только в первой хранятся данные *об условиях* правил, а во второй – *о заключениях* правил.

Сначала РТР пуста, а по мере рассмотрения правил, входящих в состав поля знаний, она пополняется новыми строками, однозначно идентифицирующимися парой «объект - атрибут объекта». Правила представляются в РТР ее столбцами. В каждую ячейку РТР записывается «тип» утверждения, он может принимать следующие значения: 0 - утверждение отсутствует в рассматриваемом правиле; 1 – утверждение присутствует в рассматриваемом правиле. Для каждого рассматриваемого правила предусмотрены два столбца: наличие утверждений в посылке и в заключении.

Применение РТР упрощает и позволяет в значительной степени автоматизировать анализ наборов правил, полученных из различных источников. Построение и анализ РТР являются лишь промежуточными этапами объединения наборов правил, полученных из различных источников. Рассмотрим подробнее основные особенности автоматизированного объединения наборов правил [19].

Для объединения двух наборов правил в единый используется анализ РТР, который сводится к подсчету совпадающих атрибутов, участвующих в правилах R_i и R_k, а также общего количества атрибутов, участвующих в данных правилах. Далее отдельно для левой и правой частей правил подсчитывается мера сходства Хемминга [16] (μ^{NL}_{ik} и μ^{NR}_{ik}):

$\mu^{N}_{ik} = n_{ik}/N$, где n_{ik} – есть число совпадающих признаков у образцов R_i и R_k,

μ^{NL}_{ik} – есть отношение количества совпавших атрибутов правых частей правил R_i и R_k к количеству всех атрибутов, участвующих в правых частях правил.

Затем формируется таблица мер схожести правил. Таблица мер схожести имеет число строк и столбцов равное суммарному числу правил, находящихся в сравниваемых наборах правил.

На первом этапе работы алгоритма создается пустая таблица, каждому столбцу и строке которой присваивается имя (номер) рассматриваемого правила. Как в столбцах, так и в строках таблицы находятся все правила, составляющие оба сравниваемых набора. На пересечении каждого столбца и строки таблицы имеется две ячейки,

Общий вид таблицы схожести правил

	R_1		…	R_N	
R_1	$\boldsymbol{\mu}^{1R}_{11}$ (1)	$\boldsymbol{\mu}^{1L}_{11}$ (1)	…	$\boldsymbol{\mu}^{NR}_{N1}$	$\boldsymbol{\mu}^{NL}_{N1}$
…	…		…	…	
R_N	$\boldsymbol{\mu}^{1R}_{1N}$	$\boldsymbol{\mu}^{1L}_{1N}$	…	$\boldsymbol{\mu}^{NR}_{NN}$ (1)	$\boldsymbol{\mu}^{NL}_{NN}$ (1)

одна предназначена для хранения меры схожести посылок, другая – для хранения меры схожести заключений. В каждую ячейку соответственно записываются правая и левая меры схожести пересекающихся правил (пересекающейся строки и столбца). Для вычисления каждой меры схожести проводится анализ РТР:

- производится выбор первой незаполненной строки таблицы мер схожести;
- в РТР выбирается столбец, номер (имя) которого равен номеру текущей строки таблицы мер схожести;
- проводится пошаговое сравнение с каждым столбцом РТР, вычисляются меры схожести посылок и заключений пар правил;
- меры схожести посылок и заключений записываются в соответствующие ячейки таблицы мер схожести;
- по окончании анализа РТР и заполнения таблицы мер схожести полученный результат сохраняется для дальнейшей обработки.

Очевидно, что главная диагональ такой таблицы будет представлена единицами, а сама таблица симметрична относительно главной диагонали, что позволяет хранить только верхнюю ее половину.

Перед началом работы процедуры объединения правил для определения последовательности вывода правил устанавливается *контрольная зона* мер схожести. В каждой строке производится анализ ячеек, содержащих соответствующие меры схожести. В случае попадания текущих меры схожести *посылки* и меры схожести *заключения* в заданный интервал, пары правил, образующие пересечение столбца и строки таблицы мер схожести помещаются в список удовлетворяющих заданным условиям и могут быть выведены для дальнейшего анализа инженером по знаниям.

По окончанию работы полученный фрагмент ПЗ подвергается завершающей обработке – все объекты и правила собираются в единый XML файл и проходят финальную перенумерацию.

Заключение

Экспериментальное программное исследование распределенного варианта КМПЗ (включающего совокупность алгоритмов и процедур совместной обработки знаний, полученных в процессе интервьюирования экспертов, анализа протоколов интервьюирования и извлечения знаний из БД) на нескольких реальных и тестовых БД показало достаточно высокую эффективность предложенного подхода как с точки зрения решения проблемы неполноты БЗ, так и поддержания БЗ в актуальном состоянии, автоматического пополнения БЗ при появлении новых БД или изменении старых БД.

В настоящее время проводится экспериментальная апробация предложенных алгоритмов и разработанных программных средств для задач медицинской диагностики, определения географического местонахождения IP-адресов, а так же задач контроля радиационных дозовых нагрузок персонала АЭС РФ.

Литература

1. Рыбина Г.В. Теория и технология построения интегрированных экспертных систем. – М: «Научтехлитиздат», 2008. - 482 с.
2. Люггер Дж. Ф. Искусственный интеллект: стратегии и методы решеня сложных проблем. – М.: Издательский Дом «Вильямс», 2003. – 864 с.
3. Осипов Г.С. Лекции по искусственному интеллекту. - М.: КРАСАНД, 2009. – 272 с.
4. Частиков А.П., Гаврилова Т.А. Белов Д.Л. Разработка экспертных систем. Среда CLIPS. – СПб.: БХВ – Петербург, 2003. – 608 с.
5. Рыбина Г.В. Основы построения интеллектуальных систем. – М.: Финансы и статистика; ИНФРА-М, 2010. - 432 с.
6. Подлипский О.К. Построение баз знаний группой экспертов // Компьютерные исследования и моделирование. 2010. Т.2. №1. с. 3-11.
7. Кобринский Б.А. Извлечение экспертных знаний: групповой вариант // Новости искусственного интеллекта. 2004. №3. с. 58-66.
8. Mendonca D., Kelton K., Rush R., Wallace W. Acquiring and Assessing Knowledge From Multiple Experts Using Graphical Representations // Knowledge-Based Systems. Vol.1. Academic Press (2000), C.T. Leondes (ed.).
9. Dieng R., Giboin A., Tourtier P., Corby O., Knowledge Acquisition for Explainable, Multi-Expert, Knowledge-Based Design Systems // EKAW. 1992. p. 298-317.
10. Feldman D., Hirsh M., Mining Associations in Text in the Presense of Background Knowledge // Proc. Of the 2nd International Conference on Knowledge Discovery (KDD-96), Portland, 1996
11. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. 2004. №3. с. 3-18.
12. Хорошевский В.Ф. Обработка естественно-языковых текстов: от моделей понимания языка к технологиям извлечения знаний // Новости искусственного интеллекта. 2002. №6. с. 19-26.
13. Quinlan J.R. Induction of Decision Trees // Machine Learning Journal. 1986. №1. p. 81-106.
14. Sreerama K., Kasif S., Salzberg S. A System for Induction of Oblique Decision Trees // Journal of Artificial Intelligence Research. 1994. №2. p. 1-32.
15. Breiman L., Fiedman J., Olshen R. and Stone C.J. Classification and Regression Trees. -Belmont, California, Wadsworth Int.Group, 1984.
16. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Издательство института математики,1999.- 210с.
17. Рыбина Г.В., Дейнеко А.О., Нистратов О.В. Особенности построения полных и непротиворечивых баз знаний в интегрированных экспертных системах // Интегрированные модели, мягкие вычисления, вероятностные системы и комплексы программ в искусственном интеллекте. Сборник научних трудов V-й международной научно-практической конференции. Т2. – М.: Физматлит. 2009. с. 760-767
18. Рыбина Г.В., Смирнов В.В. Планирование процедур верификации баз знаний в интегрированных экспертных системах // Инженерная физика. 2006. № 3. с. 53-65.
19. Дейнеко А.О., Рыбина Г.В. Распределенное приобретение знаний для автоматизированного построения баз знаний интегрированных экспертных систем // Двенадцатая национальная конференция по искусственному интеллекту с международным участием. КИИ - 2010. Труды конференции. Т.2 – М.: Физматлит, 2010. с. 240-247.

Рыбина Галина Валентиновна. Профессор кафедры кибернетики Национального исследовательского ядерного университета «МИФИ» (НИЯУ МИФИ). Окончила Московский инженерно-физический институт (государственный университет) в 1971 году. Доктор технических наук, профессор. Лауреат премии Президента РФ в области образования. Автор свыше 400 печатных работ. Область научных интересов: интеллектуальные системы и технологии, статические, динамические и интегрированные экспертные системы, интеллектуальные диалоговые системы, многоагентные системы, инструментальные средства.

Дейнеко Александр Олегович. Аспирант кафедры кибернетики Национального исследовательского ядерного университета «МИФИ» (НИЯУ МИФИ). Окончил Московский инженерно-физический институт (государственный университет) в 2008 году. Автор 8 печатных работ. Область научных интересов: интеллектуальные системы и технологии, интегрированные экспертные системы, инструментальные средства, приобретение знаний, технологии Data Mining и KDD.