ИЗВЕСТИЯ
ПЕНЗЕНСКОГО ГОСУДАРСТВЕННОГО
ПЕДАГОГИЧЕСКОГО УНИВЕРСИТЕТА
имени В. Г. БЕЛИНСКОГО
ФИЗИКО-МАТЕМАТИЧЕСКИЕ И ТЕХНИЧЕСКИЕ НАУКИ
№ 18 (22) 2010

IZVESTIA
PENZENSKOGO GOSUDARSTVENNOGO
PEDAGOGICHESKOGO UNIVERSITETA
imeni V. G. BELINSKOGO
PHYSICAL, MATHEMATICAL AND TECHNICAL SCIENCES
№ 18 (22) 2010

УДК 004.81

ПРОБЛЕМА ИЗВЛЕЧЕНИЯ ЗНАНИЙ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

© К. А. АМУРСКИЙ, В. В. ДРОЖДИН, Ю. Н. СЛЕСАРЕВ
Пензенский государственный педагогический университет им. В. Г. Белинского,
кафедра прикладной математики и информатики
e-mail: KirillAmurskii@gmail.com, drozhdin@yandex.ru, slesarevun@gmail.com

Амурский К. А., Дрождин В. В., Слесарев Ю. Н. – Проблема извлечения знаний в информационных системах // Известия ПГПУ им. В. Г. Белинского. 2010. № 18 (22). С. 96–98. – *В статье рассматривается проблема извлечения знаний на разных уровнях организации информационной системы. Дано краткое описание современного состояния данной проблемы. Проведен анализ современных методов извлечения знаний. Сделан вывод, что известные методы извлечения знаний являются недостаточными и требуют дополнительного анализа и разработки с учетом специфики информационных систем.*

Ключевые слова: самоорганизующаяся система, информационная система, самообучение, знания, методы извлечения знаний.

Amurskii K. A., Drozhdin V. V., Slesarev J. N. – The problem of knowledge extraction in information systems // Izv. Penz. gos. pedagog. univ. im.i V. G. Belinskogo. 2010. № 18 (22). P. 96–98. – *In the article the problem of knowledge extraction on different levels of the organization of information system is considered. A short description of a current state of the given problem is given. The analysis of modern methods of knowledge extraction is carried out. It is concluded that the known methods of knowledge extraction are insufficient and demand the additional analysis and working out taking into account the specificity of information systems.*

Keywords: self-organizing system, information system, self-training, knowledge, methods of knowledge extraction.

На современном этапе развития компьютерная техника достигла значительных результатов. Однако, имея огромные возможности, компьютерные системы все еще остаются пассивными, неспособными к адаптации в изменяющейся среде и самосохранению. Следствием пассивности являются, с одной стороны, необходимость сложного и трудоемкого процесса проектирования и создания систем, а, с другой – их быстрое моральное старение и прекращение использования.

Необходимым условием приобретения активности являются накопление информации, извлечение знаний и использование их в функционировании системы. Активность системы позволит запустить процессы самоорганизации, являющиеся самыми мощными механизмами адаптации систем.

Создание реальных самоорганизующихся систем целесообразно начать с совершенствования автоматизированных информационных систем.

СВОЙСТВА САМООРГАНИЗУЮЩИХСЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Самоорганизующейся информационной системой является система, способная активно поддерживать свое существование и обеспечивать решение информационных задач с требуемым качеством в течение длительного (потенциально бесконечного) времени в условиях существенных изменений внешней среды и внутренней организации системы. Для обеспечения такого качества информационная система должна обладать следующими свойствами:

– обеспечение организации и обработки данных в рамках эволюционной модели данных [1, 2];

– открытость системы на всех уровнях организации, для обеспечения высокой адаптивности к изменениям внешней среды и внутренней организации системы [2, 3];

– реализация принципа структурно-функционального единства системы, предполагающего очень высокий уровень адекватности (согласованности) структурной и функциональной организаций системы на основе их оптимизации, что способствует приобретению свойств интенсивных систем (функционирование в ограниченном объеме с возрастанием организованности и сложности в процессе существования системы);

– активность системы, базирующаяся на способности самостоятельного принятия решений и выполнении различных действий на основе своих внутренних потребностей, но в интересах внешней среды, создает условия повышения ее полезности для внешней среды и более длительного существования;

– поддержка уровня организованности системы, достаточного для противодействия агрессивности внешней среде;

– взаимодействие с внешней средой как с системой, определенным образом организованных объектов (пользователей), способствует повышению интенсивности их взаимодействия и возникновению процесса коэволюции, существенно ускоряющего темпы их развития, и созданию на этой основе принципиально новой системы более высокого иерархического уровня – сверхсистемы [3].

ПРОБЛЕМА САМООБУЧЕНИЯ

Самообучение заключается в формировании собственных индивидуальных знаний (опыта) и использования их в своем функционировании.

Известно, что знания бывают двух типов: эмпирические и теоретические. Эмпирические знания являются некоторой аппроксимацией известных фактов, которые конкретный объект воспринимал в процессе своего существования, а теоретические знания являются абстракциями, отражающими предельные случаи или сформированные теоретическим путем на основе аксиоматического подхода.

Учитывая, что самоорганизующихся информационных систем пока не существует, то целесообразно решать задачу самообучения на достаточно простом уровне – выявление и использование системы эмпирических знаний.

Для решения этой проблемы информационная система обладает рядом преимуществ:

– она содержит модель мира, которая может быть расширена информацией о самой себе;

– выполняет ограниченный набор действий (сбор, хранение, обработку и выдачу информации) и при этом является основой для создания многих других систем;

– обладает большими возможностями для модификации и адаптации на уровнях организации данных, обработки запросов, взаимодействия с внешней средой и др.;

– имеет хорошую формализацию на основе теории баз данных и языков запросов.

Это создает высокую определенность в функционировании системы и предоставляет большой объем информации (фактов) для выявления различных закономерностей, что позволяет сформировать целостную надежную систему эмпирических знаний и использовать ее в функционировании системы. По сути, самообучение представляет собой порождение в системе внутренней семантики и включение ее в структуру системы. Таким образом, самообучение – это механизм реализации принципа структурно-функционального единства системы, являющегося базовым принципом существования живых систем.

Большие возможности современной компьютерной техники (большие объемы памяти и высокое быстродействие, локальные и глобальные сети) и достижения в области создания программного обеспечения (многоплатформенность, библиотеки времени исполнения, языки, ориентированные на формальное описание предметной области (DSL), методы извлечения знаний из баз данных и др.) создают принципиальную основу для решения проблемы самообучения.

Однако при самообучении возникает две серьезных проблемы:

– как определить, что является существенным, а что второстепенным для функционирования системы;

– все алгоритмы обучения имеют очень высокую временную сложность (являются NP-полными), т. к. базируются на больших переборах вариантов.

УРОВНИ ВЫЯВЛЕНИЯ ЗАКОНОМЕРНОСТЕЙ

В информационной системе можно выделить три уровня выявления закономерностей:

а) на уровне организации данных;

б) на уровне обработки запросов (последовательности и совместная обработка данных);

в) на уровне использования данных пользователями.

Целью формирования системы эмпирических знаний на уровне организации данных является повышение надежности (устойчивости) и эффективности обработки данных путем настройки адаптивных структур данных и методов их обработки на реализацию требуемых запросов.

Для достижения этой цели необходимо решение следующих задач:

– выявление факторов, существенно влияющих на организацию данных;

– выявление наиболее сильных (логических, функциональных) зависимостей между данными;

– разработка критериев определения допустимых отклонений от логической организации данных в форме исключений;

– разработка метода поддержки непротиворечивости и целостности на уровне организации данных.

Целью формирования системы эмпирических знаний на уровне обработки запросов является повышение надежности (устойчивости) и эффективности обработки данных путем оптимизации запросов и учета закономерностей последовательной и совместной обработки данных.

Для достижения этой цели необходимо решение следующих задач:

– выявление факторов, существенно влияющих на реализацию запросов;

– выявление зависимостей между данными и использующими их запросами;

– разработка критериев определения допустимой обработки данных;

– разработка метода поддержки корректной обработки данных.

Целью формирования системы эмпирических знаний на уровне использования данных пользователями является повышение надежности (устойчивости) и эффективности обработки данных путем формирования задач и построения функциональных моделей пользователей.

Для достижения этой цели необходимо решение следующих задач:

– выявление факторов, существенно влияющих на решение задач пользователями с помощью информационной системы;

– выявление устойчивых последовательностей запросов от различных пользователей и формирование информационных задач;

– разработка критериев определения допустимых задач, решаемых пользователями;

– разработка метода поддержки корректного решения информационных задач пользователями.

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ЗНАНИЙ В СУЩЕСТВУЮЩИХ СИСТЕМАХ

В существующих системах используются следующие подходы к извлечению эмпирических знаний: статистические методы, нейросетевые алгоритмы, деревья решений, алгоритмы ограниченного перебора, методы рассуждений на основе аналогий, генетические алгоритмы, системы визуализации многомерных данных и методы нечеткой логики [4, 5].

Наиболее часто для извлечения знаний в различных системах используются статистические методы. Эти методы обладают высокой универсальностью и хорошо разработаны, но требуют репрезентативных выборок и обладают низкой информативностью. Оценки, полученные на основе статистических методов, являются нижними границами для семантических методов.

Нейронные сети тоже очень широко распространены. При всей их привлекательности они имеют ряд существенных недостатков: для обучения требуется большой обучающий набор, возможен эффект переобучения, но самым важным недостатком является неинтерпретированность сформированной системы знаний. Вследствие этих недостатков нейронные сети имеют ограниченное применение.

Ряд задач, связанных с извлечением знаний, эффективно решается с помощью деревьев решений. Деревья решений – это способ представления правил в виде иерархической структуры, в которой каждой ситуации соответствует единственный узел, дающий решение. Под правилом понимается продукция вида «если x, то y». С помощью деревьев решений обычно решают задачи описания и классификации объектов, регрессия и др. Однако большинство известных алгоритмов формирования и обработки деревьев решений являются «жадными алгоритмами», поэтому эффективность их использования достаточно низка.

Методы рассуждений на основе аналогий для прогнозирования и принятия решений ищут в прошлом близкие аналоги ситуации и выбирают ответ, который был для них правильным. Недостатками таких методов являются сложность построения моделей, обобщающих предыдущий опыт, что приводит к произвольному выбору «мер близости» ситуаций.

Генетические алгоритмы – адаптивные методы поиска, часто используемые для решения задач функциональной оптимизации. Основным преимуществом таких алгоритмов является их способность манипулировать одновременно многими параметрами. Недостатками является то, что они не могут эффективно применяться для небольшого пространства поиска вследствие высокой вероятности схождения к локальному оптимуму, а не к глобально лучшему решению, а также высокая зависимость эффективности от методов кодирования решений, операторов настройки параметров, частных критериев успеха.

Математическая теория нечетких множеств и нечеткая логика, являющиеся обобщениями классической теории множеств и формальной логики, предназначены для построения нечетких и приближенных рассуждений при описании человеком процессов и систем. Однако высокая неопределенность в построении функций принадлежности и формирование множественного ответа существенно снижают качество нечеткого вывода.

Таким образом, известные методы извлечения знаний являются недостаточными и требуют дополнительного анализа и разработки с учетом специфики информационных систем. Поэтому целесообразна разработка методов извлечения знаний, ориентированных на:

– использование на всех уровнях организации информационной системы;

– формирование четкой системы зависимостей оптимальной сложности [6];

– расширение системы зависимостей исключениями, обеспечивающими ее надежность и совершенствование;

– возможность выявления зависимостей независимо от специфики наборов данных;

– формирование устойчивой системы эмпирических знаний, описывающей предметную область с требуемой точностью;

– эффективную обработку данных в процессе решения информационных задач.

СПИСОК ЛИТЕРАТУРЫ

1. Дрождин В. В. Системный подход к построению модели данных эволюционных баз данных // Программные продукты и системы. 2007. № 3 С. 52–55.
2. Дрождин В. В. Открытость структур в эволюционной модели данных // Программные продукты и системы. 2009. № 2. С. 135–137.
3. Дрождин В. В., Зинченко Р. Е. Системный подход к концептуальному моделированию предметной области в самоорганизующейся информационной системе // Программные продукты и системы. 2009. № 4. С. 73–79.
4. Дюк В. А., Самойленко А. П. Data Mining: учебный курс. СПб: Питер, 2001. 368 с.
5. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. М., 2004.
6. Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. М., 1982. 296 с.