УДК 004.896

Д.Е. Семёнов

Модификация модели представления информационных объектов с использованием ориентированных графов в реляционных базах данных*

Рассматриваются и исследуются существующие модели представления данных в виде онтологий. Описана структура реляционной базы данных для модифицированной модели информационных объектов, построенная в виде ориентированных графов.

Ключевые слова: онтология, ориентированный граф, семантический поиск, интеллектуальные системы, реляционные базы дынных.

Введение

В связи с большими объемами накопленной информации в электронном виде, возникает потребность структурированного хранения и поиска необходимых данных, что является весьма актуальной задачей, особенно это связано с лавинообразным развитием информационных технологий. Исследования в этой области являются перспективным направлением, об этом свидетельствует большой объем опубликованных материалов.

В настоящее время большое развитие получило направление Semantic Web, предложенная консорциумом W3C [1], целью которого является реализация возможности машинной обработки информации. Основной акцент данной концепции основывается на работе с метаданными, однозначно характеризующими свойства и содержание электронных документов, построенных в виде информационных объектов (ИО). Принцип построения таких документов представлен в работе [2] и может быть рассмотрен в трех разных аспектах – структура, контент и контекст. Для описания семантических моделей (онтологий) появились стандарты (например, RDF, OWL, SPARQL и др.), а также множество инструментов для работы, таких как редакторы онтологий (например, Protégé) и системы логического вывода (например, Pellet, Racer, Joseki и т.д.) [1]. Семантический подход (информационная модель предметной области, представляющая собой ориентированный граф) может быть использован для обеспечения поиска необходимых данных в профессионально-ориентированных информационных системах, которые стали активно появляться на рынке программных продуктов еще в 60-х годах прошлого века [5]. Существующие модели описания ИО, построенные на ориентированных графах, представляют собой реляционное отношение множества пронумерованных вершин (документов) и характера связи между ними [2, 3]. В работе [4] описана методология использования Semantic Web в организациях для решения задачи интеграции разнородных ресурсов информации и данных.

Одним из важнейших этапов хранения информации является классификация документов, поскольку отсутствие классифицированных признаков ведет к увеличению времени поиска. При отсутствии такой классификации последующая обработка интеллектуальными информационными системами (ИИС) может сводиться к простому полнотекстовому поиску, при использовании такой технологии время обработки запроса значительно увеличивается.

Целью данной работы является модификация существующей модели представления электронных документов, за счет оптимизации структуры хранения информации в реляционных базах данных. Для достижения поставленной цели необходимо: провести анализ существующих моделей представления электронных документов; модифицировать способ хранения информации, обеспечивающий компактное хранение и быстрый поиск; сформировать структуру таблиц реляционной базы данных.

Основные понятия

Для работы с неструктурированным набором информационных объектов и эффективного семантического поиска предполагается использовать онтологическое моделирование. Онтология предметной области включает спецификации понятий и отношений между ними. Модель онтологии в данной статье формально может быть представлена как множество

* Работа выполнена при финансовой поддержке гранта «Индивидуальный грант молодого ученого Томского политехнического университета», приказ ректора от 21.05.2010 г. № 3401.

$$O = \{X, R, F\}, \quad (1)$$

где, X – конечное множество понятий предметной области; R – конечное множество отношений между понятиями; F – конечное множество функций интерпретации.

В предлагаемом подходе становится возможным представить информационный объект (документ) в виде онтологии. Такой документ имеет обязательный заголовок, содержательную часть, ключевые слова, метаописания. Дополнительные атрибуты могут представлять собой ссылки на другие информационные объекты или дополнительные тестовые поля. Все документы могут иметь множественные перекрёстные ссылки. К таким документам можно отнести персональную карточку сотрудника или текстовый вопрос в тестовой форме с различными вариантами ответа для компьютерных программ тестирования, хранящиеся в одном или разных электронных документах, т.е. данные, построенные в виде иерархий.

Модификация модели представления информационных объектов, построенных на ориентированных графах

Сравнивая и анализируя подходы представления ИО, описанные в работах Е.А. Сидоровой [2], А.В. Маликова [3], принят во внимание тот факт, что однотипные документы никак не формируются в группы, а имеют ссылки, объединяющие их в единое подмножество. Однотипными документами будем считать информационные объекты, имеющие одинаковые наборы атрибутов. Для выделения схожих, но не связанных между собой документов нет критериев, по которым они могут быть объединены в единое множество; различные атрибуты, значения которых указывает на другие ИО, никак не отличаются друг от друга.

В работе [3] (концепция, в виде структуры СУБД приведена в табл. 1) для идентификации ИО используются натуральные числа, что может привести к конфликтной ситуации при объединении двух и более множеств (1) электронных документов. Этот факт связан с тем, что при слиянии разных ИО в единое хранилище возникнет дублирование первичных ключей идентификатора вершины (это видно из предложенной структуры).

Таблица 1

Концепция типа «предки – потомки», описанная в работе А.В. Маликова

Таблица с данными о вершинах: CREATE TABLE HierarchyObject (ID int NOT NULL primary key, name varchar(50), level int)	 – идентификатор вершины, – наименование вершины, – уровень вершины в ориентированной сети
Таблица с данными о достижимости между вершинами: CREATE TABLE HierarchyLink (IDancestor int NOT NULL, IDdescendant int NOT NULL)	 – идентификатор вершины предка, – идентификатор вершины потомка

В результате исследований предложено использовать стандарт идентификации Universally Unique Identifier (UUID) [5], предназначенный для создания программного обеспечения. Все используемые идентификаторы (ID) ИО предлагается преобразовать в глобальные статистически уникальные 128-битные номера. В шестнадцатеричной системе счисления UUID выглядит как: «12345678-1234-1234-1234-1234567890AB». Введение нового типа идентификации позволит объединять различные множества $O_{all} = \{O_1, O_2, \ldots, O_n\}$ без возникновения конфликтных ситуаций, вызванных совпадением идентификаторов. После изменения типа идентификатора с INT на CHAR(32) произойдет увеличение размера таблиц, так как для хранения 128-битных номеров, представленные в шестнадцатеричном формате, требуется в 8 раз больше места, чем для натуральных чисел. Хотя уникальность каждого отдельного ID не гарантируется, общее количество уникальных ключей настолько велико (2^{128}), что вероятность того, что в мире будут независимо сгенерированы два совпадающих ключа, достаточно мала.

Модификация модели заключается в следующем: добавление ID типа ИО, позволяющего группировать несвязанные однотипные объекты в отдельные множества; идентификатор объектов сделать также уникальным, а не целочисленным, как в существующей модели А.В. Маликова [3]; все атрибуты, являющиеся ссылками на другие ИО, объединять в группы по концепциям и каждой группе установить ID.

Таким образом, после модификации представления документов становится возможным модель ИО представить в следующем формальном виде:

$$O_{\text{mod}} = \{X', R', F, F_{tag}\}, \quad (2)$$

где, X' – конечное множество понятий предметной области с уникальным ID; R' – конечное множество отношений между понятиями, сгруппированное по концепциям; F – конечное множество функций интерпретации; F_{tag} – конечное множество функций расстановки ключевых слов к ИО.

Взаимосвязь этих документов представляется возможным записать в виде ориентированного графа (пример орграфа представлен на рис. 1), вершинами которого являются электронные документы, а дугами – связи между ними.

$$G = \{V, A\}, \quad (3)$$

где, V – множество вершин или узлов; A – множество (упорядоченных) пар различных вершин, называемых дугами или ориентированными рёбрами.

У каждой дуги свой собственный вес, значение которого соответствует идентификатору связанного объекта. Каждый документ имеет идентификаторы ID, указывающие на объекты, с которыми он связан, и ID объектов, ссылающихся на него. Таким образом, осуществляется двухсторонняя связь, проиллюстрированная на рис. 1.

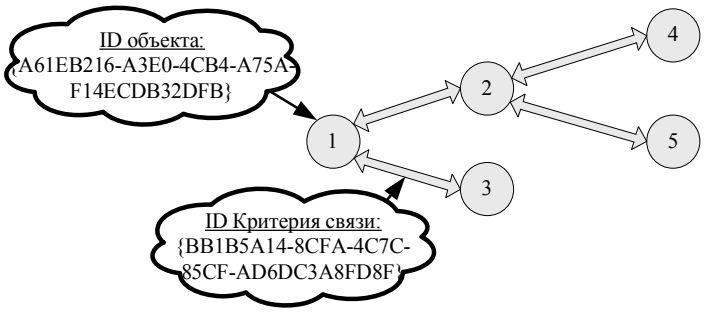

Рис. 1. Пример взаимосвязи ИО в виде ориентированного графа

Практическая реализация

Моделируя предложенную структуру представления ИО (2), становится возможным составить матрицу достижимости (бинарная матрица замыкания), описывающую информацию о существовании путей между вершинами ориентированного графа. Все вершины орграфа полагаются равнозначными и могут иметь неограниченное количество уникальных перекрестных ссылок. Оперировать с такой матрицей становится возможным при использовании реляционной модели хранения данных, реализованной в виде системы управления базами данных (СУБД), базирующихся на теории множеств. СУБД представляет собой совокупность специальных языковых и программных средств, облегчающих пользователям выполнение всех операций, связанных с организацией хранения данных, их корректировкой и доступом к ним. В этом случае матрица достижимости представляется ограниченным набором картежей одной таблицы в БД.

Для обеспечения хранения информации о матрице достижимостей вершин графа (3) в СУБД необходимо составить одну таблицу «tbl_Links», состоящую из трех столбцов: первый столбец содержит информацию об идентификаторе одного информационного объекта, второй столбец – идентификатор другого документа, третий – идентификатор концепции связи (пример структуры приведён в табл. 2). Все столбцы являются обязательными для заполнения, так как указывают на связь между двумя вершинами орграфа по конкретному критерию.

Таблица 2

Пример содержания бинарной матрицы замыкания вершин ориентированного графа

ID	LINK_ID	LINK_TYPE
{B2312809-1B8F-4A96-9505-5D752C83C77D}	{35B34436-4FB6-4147-8A5B-634562C3AD52}	{5F6938B8-31D2-4B23-A713-D3F604D7DA2F}
{35B34436-4FB6-4147-8A5B-634562C3AD52}	{35B34436-4FB6-4147-8A5B-634562C3AD52}	{5F6938B8-31D2-4B23-A713-D3F604D7DA2F}

Информация об ИО сохраняется в таблице «tbl_Objects» (пример структуры приведён в табл. 3), в которой присутствуют обязательные поля: ID объекта; ID типа ИО; заголовок объекта; метаданные, необходимые для обеспечения семантического поиска, и дру-

гие поля, необходимые для описания ИО (время создания; ID пользователя, создавшего ИО; атрибуты и др.).

Таблица 3

Пример содержания таблицы с информационными объектами

ID	CLASS_ID	Caption	...	Meta_description
{B2312809-1B8F-4A96-9505-5D752C83C77D}	{C18DD665-B5FF-44BA-BB9A-F7CEF06A101F}	Имя документа 1		Описание объекта 1
{35B34436-4FB6-4147-8A5B-634562C3AD52}	{C18DD665-B5FF-44BA-BB9A-F7CEF06A101F}	Имя документа 2		Описание объекта 2

Любой ИО может содержать неограниченное множество дополнительных атрибутов, хранение такой информации достигается за счет введения двух дополнительных таблиц. В первой таблице «tbl_Keys» (пример представлен в табл. 4) содержится описание ключей для конкретного класса документов. В этой таблице содержатся следующие поля: ID атрибута; ID типа ИО, к которому принадлежит это поле; описание, значение по умолчанию; признак того, что поле является обязательным для заполнения; признак того, что поле является ссылкой.

Таблица 4

Пример таблицы описания дополнительных полей ИО

FIELD_ID	CLASS_ID	Description	Is_Link	Required
{F2057214-9899-4D65-8133-C6A8C9A60D6E}	{C18DD665-B5FF-44BA-BB9A-F7CEF06A101F}	Фамилия	False	True
{7CA34328-A959-4EB0-9098-9A672BCD5736}	{C18DD665-B5FF-44BA-BB9A-F7CEF06A101F}	Имя	False	True
{09917887-3E2F-4281-8FCD-27F7B9B07EA7}	{ACD32778-B395-4678-A0C3-F830A7AD038F}	Отчество	False	True

В следующей таблице «tbl_Values» (пример представлен в табл. 5) содержится значение конкретного атрибута для определенного информационного объекта. Здесь содержится информация об ID объекта, ID поля и его значение.

Таблица 5

Пример таблицы с содержанием значений атрибутов ИО

FIELD_ID	ID	Value
{F2057214-9899-4D65-8133-C6A8C9A60D6E}	{B2312809-1B8F-4A96-9505-5D752C83C77D}	Иванов
{7CA34328-A959-4EB0-9098-9A672BCD5736}	{B2312809-1B8F-4A96-9505-5D752C83C77D}	Иван
{09917887-3E2F-4281-8FCD-27F7B9B07EA7}	{B2312809-1B8F-4A96-9505-5D752C83C77D}	Иванович

Для обеспечения хранения ключевых слов ИО, необходимых для семантического поиска, предложена таблица «tbl_Tags» (пример представлен в табл. 6), содержащая ID объекта и ключевое слово, с ним связанное. За установкой такого отношения отвечает пользователь, работающий с документом. Чем больше одинаковых слов встречается в таблице, тем больше их вес, и наоборот, чем реже встречается слово – тем меньше вес.

Таблица 6

Пример таблицы, содержащей ключевые слова документов

ID	Value
{F2057214-9899-4D65-8133-C6A8C9A60D6E}	Преподаватель
{7CA34328-A959-4EB0-9098-9A672BCD5736}	Научный руководитель
{09917887-3E2F-4281-8FCD-27F7B9B07EA7}	Победитель конкурса

Такой подход при проектировании реляционной базы данных позволяет хранить большое количество разнотипных документов в существующих таблицах. Добавление нового атрибута информационного объекта не повлечет изменение структуры БД, что существенно позволяет сократить время разработки программного продукта, не зависящего от конкретной предметной области.

Приведённое текстовое описание структуры реляционной БД может быть представлено согласно методологии функционального моделирования (Function Modeling) с помощью графической нотации IDEF0 в виде набора функциональных блоков (рис. 2), показывающих иерархию взаимосвязей таблиц.

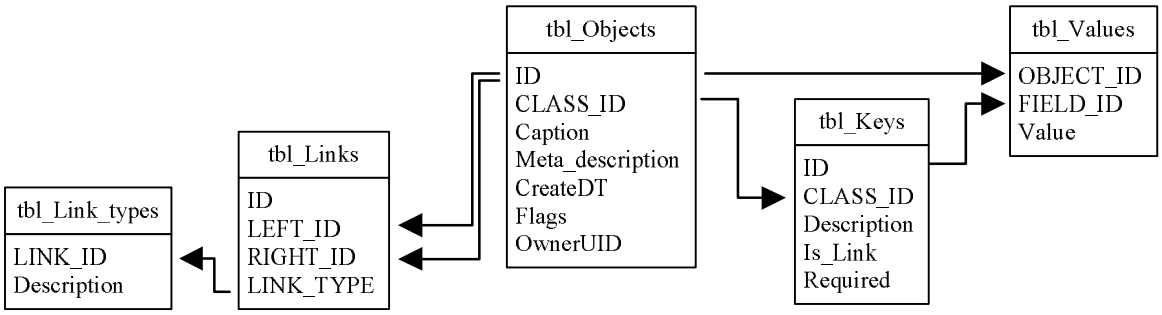

Рис. 2. Графическое представление структуры реляционной базы данных

На базе предложенной модели представления информационных объектов реализованы программные компоненты в среде быстрой разработки приложений Embarcadero RAD Studio 2009, выполненные в виде COM-объектов. Хранение данных реализовано в реляционной СУБД, поддерживаемой на данный момент, – MySQL и Microsoft® SQL Server. Таким образом, предложенный процесс реализации ИО относится к объектно-ориентированной методологии разработки программных продуктов, являющейся приоритетной в настоящее время.

Применение на практике

Рассмотрим несколько примеров манипулирования данными, демонстрирующих особенности предложенной модифицированной модели представления информационных объектов. Запросы построены с использованием языка Structured Query Language (SQL), приведены для предложенной структуры СУБД:

Задача 1: загрузить список сотрудников кафедры, а именно найти все идентификаторы и заголовки объектов одного класса @Class_ID1, сформировать множество однотипных объектов. Запрос на языке SQL имеет вид

SELECT ID, CAPTION FROM tbl_Objects WHERE CLASS_ID=@Class_ID1;

Задача 2: получить анкету сотрудника, а именно загрузить документы @ID1 с полным набором атрибутов:

SELECT * FROM tbl_Objects WHERE ID= @ID1;

Предложенный SELECT формирует базовую информацию об одном информационном объекте, включая тип объекта (@Class_ID1), которая будет использоваться далее.

SELECT tbl_Values.Value, tbl_Keys.Description FROM tbl_Keys LEFT JOIN tbl_Values USING FIELD_ID WHERE tbl_Keys.CLASS_ID=@Class_ID1 AND tbl_Values.ID=@ID1;

В результате выполнения этого запроса формируется таблица, содержащая все значения атрибутов указанного информационного объекта. Для достижения требуемого результата за меньшее количество обращений к СУБД возможно объединение SQL инструкций в единый запрос.

Для всех атрибутов отношений, участвующих в SQL запросах, строятся индексы, что позволяет добиваться высокой производительности. Большая скорость извлечения данных из СУБД осуществляется за счет использования простой структуры таблицы с обозначенными бинарными ключами.

Пример использования

Рассмотрим пример расположения двух электронных документов, содержащих информацию о сотрудниках одной организации, в предложенной концепции построения информационных объектов. На рис. 3 эти сотрудники обозначены: «Разработчик 1» и «Разработчик 4» (на рис. 3 – темный круг). В каждом документе содержится информация о фамилии, имени, отчестве, должности, месте работы (на рис. 3 – светлый круг) и т.д. Сплошной тонкой стрелочкой обозначена прямая связь между информационным объектом и атрибутами ИО, жирной стрелочкой обозначена связь между двумя документа-

ми. Пунктирной линией обозначены документы, связанные с ИО. В «облаке» указаны коды идентификаторов, связывающие ИО.

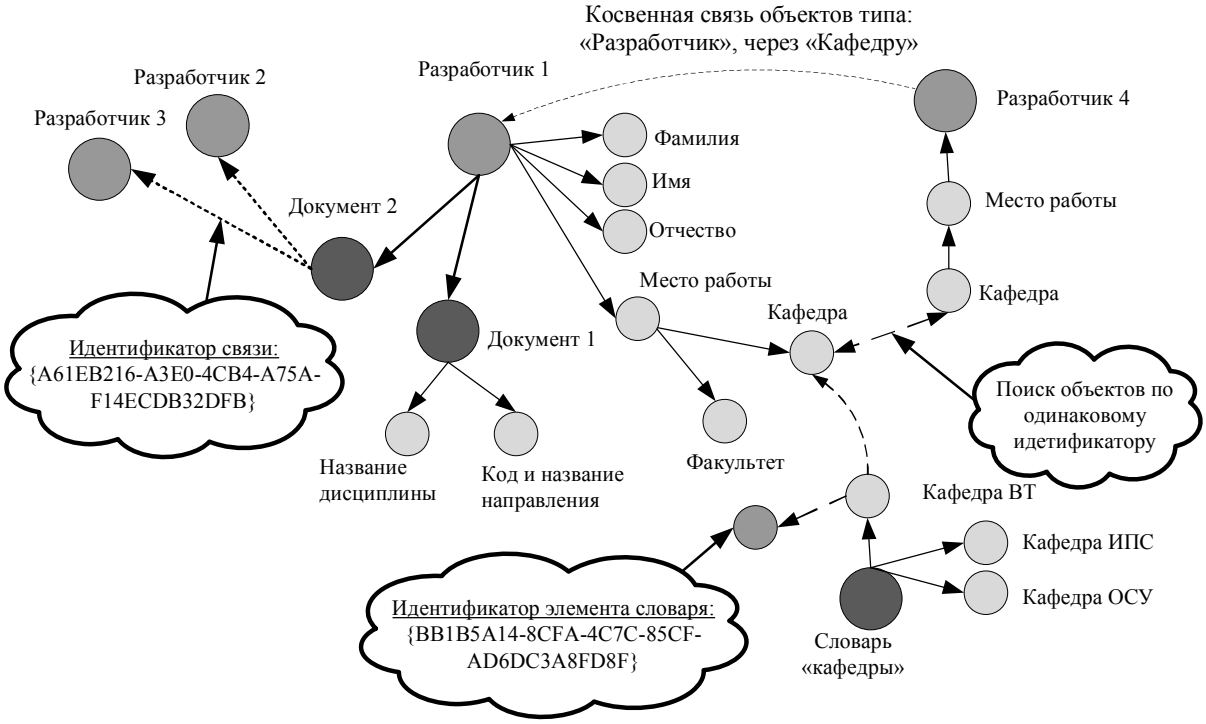

Рис. 3. Пример построения информационного пространства с использованием информационных объектов

Все изменяющиеся атрибуты, такие как место работы, список читаемых дисциплин загружаются из редактируемых словарей (множество однотипных информационных объектов определенного класса). За наполнение и поддержание электронных словарей в актуальном состоянии отвечает администратор информационной системы. У информационного объекта, представляющего информацию о разработчике, есть атрибут «кафедра», находящийся у объекта «место работы». Значение этого атрибута заполняется из существующих словарей-документов, отражающих полный список кафедр в учебном заведении, и содержит уникальный ID кафедры. Далее может быть проведено сопоставление с другим сотрудником, у которого в атрибуте «кафедра» установлен идентичный идентификатор. Таким образом, два вершины в орграфе связаны друг с другом через третью вершину.

Выводы

Новизной в данной работе является введение уникальных статических номеров Universally Unique Identifier, позволяющих использовать специальные идентификаторы преднамеренно, для повторной идентификации той же самой сущности в различных контекстах. Стандарт UUID обеспечивает исключения конфликтных ситуаций, вызванных совпадением идентификаторов при объединении множеств информационных объектов. Предложена структура реляционной базы данных, предназначенная для хранения бинарной матрицы замыкания и информации о вершинах ориентированного графа, а также дополнительной информации об ИО. Эта структура может быть использована при создании схемы БД. Представлены запросы получения результатов выборки данных из реляционной СУБД.

В настоящее время предложенный в этой статье подход для организации хранения, обработки, поиска информации используется в основе интеллектуальной информационной системы экспертизы качества тестовых материалов [6], которая направлена на проведение содержательной (предметной) и тестологической экспертизы тестовых материалов, позволяющих минимизировать ошибку измерения уровня учебных достижений учащихся, а также предоставит преподавателю надежный и объективный дидактометрический инструмент.

Литература

1. Allemang D. Semantic Web for the Working Ontologist: Modeling in RDF, RDFS and OWL / D. Allemang, J. Hendler. – Morgan Kaufmann Publishers, 2008. – 350 p.
2. Сидорова Е.А. Семантический подход к анализу документов на основе онтологии предметной области / Е.А. Сидорова, Ю.А. Загорулько, И.С. Кононенк // Компьютерная лингвистика и интеллектуальные технологии: труды междунар. конф. «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2006. – С. 468–474.
3. Маликов А.В. Ориентированные графы в реляционных базах данных // Докл. Том. гос. ун-та систем управления и радиоэлектроники. – 2008. – № 2(18), ч. 2. – С. 100–104.
4. Черний А.В. Развитие информационной системы организации с использованием семантических технологий / А.В. Черний, А.Ф. Тузовский // Матер. всерос. конф. с междунар. участием «Знания–Онтологии–Теория» (Новосибирск, 20–22 октября 2009 г.). – Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. – Т. 2. – С. 52–59.
5. Электронная энциклопедия «Wiki» [Электронный ресурс] – Режим доступа http://ru.wikipedia.org/wiki/, свободный (дата обращения 10.03.2010).
6. Муратова Е.А. Информационная система поддержки экспертизы качества тестовых материалов / Е.А. Муратова, Д.Е. Семёнов // «Интеллектуальные системы» (AIS-08) и «Интеллектуальные САПР» (CAD-2008): труды междунар. науч.-техн. конф. (Дивноморское, 3–10 сентября 2008 г.): В 4 т. – М.: Физматлит, 2008. – Т. 1. – С. 182–190.

Семёнов Дмитрий Евгеньевич
Аспирант каф. инженерной педагогики ИДНО ТПУ
Тел.: (8-382-2) 56-40-82
Эл. почта: dimomans@tpu.ru

Semenov D.E.

Modification of the model of data entity presentation using directed graphs in relational databases

The paper presents and investigates existing models of data entity presentation in the form of ontology. A modified model for the description of relational databases designed in the form of directed graphs has been suggested.

Keywords: ontology, directed graphs, relational databases, Semantic Web, Intellectual system.