УДК 004.358:681.518

Л.А. Пономаренко, В.А. Филатов, А.Ю. Шевякова

ОБ ОДНОМ ПОДХОДЕ К РЕШЕНИЮ ЗАДАЧИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ДАННЫХ

Статья посвящена исследованию и разработке эффективных моделей и методов интеллектуального анализа данных в информационных системах.

Введение

Развитие средств вычислительной техники и увеличение объемов хранимой информации привело к необходимости выделения технологии баз данных в отдельную область интенсивных научных исследований. За более чем 40 летнюю историю базы данных прошли развитие от автоматизированных мест, так называемых АРМов, технологий файл и клиент сервер до информационных пространств. Управление информацией претерпевает изменения в ряде его аспектов: в моделях хранения и доступа, в масштабах проектируемых систем – от баз данных до систем поддержки принятия решений. Исследованию вопросов, связанных с современным представлением информации, информационных ресурсов и информационных технологий посвящено ряд работ [1].

Рассматриваемая статья посвящена исследованию методов и моделей построения информационных систем, объединяющих в себе реляционные базы данных как источник первичных данных и средства интеллектуального анализа.

Целью работы является исследование и разработка эффективных моделей и методов интеллектуального анализа данных информационных систем.

2. Реляционная база данных, как среда хранения данных в интеллектуальных аналитический системах

Реляционная база данных - база данных, основанная на реляционной модели. Слово «реляционная» происходит от английского relation - отношение. Эта модель характеризуется простотой структуры данных, удобным для пользователя табличным представлением и возможностью использования формального аппарата алгебры отношений и реляционного исчисления для обработки данных.

Для построения структурной схемы баз данных используются традиционные средства спецификации реляционной модели данных. Основной структурной единицей данных в реляционной модели является *n*-арное отношение, представляющее собой конечное подмножество декартова произведения доменов, т.е множеств атомарных значений элементов данных – атрибутов отношения.

Пусть R – конечное множество имен отношений базы данных;

D – $D_1,...,D_i$ – множество доменов, где всякий домен D_i есть именованное множество атомарных значений элементов данных;

A – конечное множество имен атрибутов отношений;

dom – отображение из A в D, определяющее из какого домена выбираются значения атрибутов.

Пару $\langle A_i, domA_i \rangle$, где $A_i \in A$ называют атрибутом.

Структурную схему S_i отношения R_i $R_i \in R$ можно представить в виде $R_i(A_1,..., A_n)$, в котором все A_i различны. Отношение r_i можно определить как расширение схемы $S_i : r_i \subseteq domA_1 \times ... \times domA_n$.

Перестановка атрибутов в схеме не порождает нового расширения и множество $A_1,...A_n$ атрибутов отношения R_i задает тип отношения. Для спецификации состава носителя используется выражение $R_i = A_1...A_n$. Структурная схема *U* реляционной базы данных – это спецификация вида $R(R_1,..., R_p)$, где $R_i \in R$ и все R_i различны.

Рассматривается задача обнаружения знаний в реляционных базах данных, эффективное решение которой, позволит повысить качество принимаемых управленческих решений.

3. Разработка метода извлечения знаний из данных представленных реляционной моделью

Формально задача извлечения знаний из базы данных может быть представлена в следующем виде. Предметная область отображается в виде реляционной модели, которая описывается универсальным отношением R, являющимся подмножеством кортежей декартового произведения $R = (DX_1, DX_2,..., DX_n, DY_1,..., DY_m) = < x_1,..., x_n, y_1,..., y_m >$ где x_i – значения входных атрибутов X_i из домена DX_i; y_i – значения выходных атрибутов Y_j из домена DY_j; $P(x_1,..., x_n, y_1,..., y_m)$ – предикат, описывающий условия отображения конкретной предметной области в кортежи значений атрибутов $< x_1,..., x_n, y_1,..., y_m >$.

Целью поставленной в исследовании задачи является формирование отображения в виде набора правил: $\{X_1,.X_2,..., X_n\} \to \{Y_1, Y_2,..., Y_m\}$ ставящих каждому входному набору значений $\{x_i = DX_i, i = \overline{1,n}\}$, в соответствие некоторый набор целевых значений $\{y_j = DY_j, j = \overline{1,m}\}$.

Полученные функциональные зависимости $Y_j = F_j(X_1, X_2, ..., X_n), j = \overline{1,m}$ должны быть верны для кортежей отношения и могут быть использованы при нахождении выходных атрибутов Y_j для новых значений входных атрибутов $X_i, i = \overline{1,n}$.

Если база данных представлена в виде сильно типизированного отношения, например в третьей нормальной форме, то для последующих расчетов необходимо перейти к универсальному отношению. Для преобразования модели можно воспользоваться операцией соединения, входящей в основные операции реляционной алгеброй [2].

Соединение эквивалентно следующей последовательности реляционных операций:

- переименовать одинаковые атрибуты в отношениях A и B;
- выполнить операцию декартово произведение отношений A и B;
- выполнить операцию выборки по совпадающим значениям атрибутов, имевших одинаковые имена отношений A и B;
- выполнить операцию проекции, удалив повторяющиеся атрибуты из отношений A и B;
- переименовать атрибуты отношений A и B, вернув им первоначальные имена.

В общем случае полученное таким преобразованием универсальное отношение – это набор фактов, каждый из которых описывается некоторым конечным набором дискретных атрибутов.

Для дальнейшего анализа данных может быть применен один из методов Data Mining – метод построения дерева решений (Decision Trees). Деревья решений являются самым распространенным в настоящее время подходом к выявлению и отображению логических закономерностей в данных. Среди них:

- ID3 (Interactive Dichotomizer)
- CART (classification and regression trees)
- CHAID(chi square automatic interaction detection)

Рассмотрим подробно процесс построения деревьев решений на примере системы ID3.

Алгоритм генерации дерева решений ID3 – это алгоритм, который строит дерево от корня к листьям, в каждом узле выбирая атрибут, который наилучшим образом классифицирует данные [3].

Входные параметры алгоритма: Examples – текущий обучающий пример - целевой атрибут, Attributes – множество атрибутов-кандидатов.

Рассмотрим общую схему работы алгоритма на примере универсального отношения, представленного в таблице 1.

Выбираем целевой атрибут - это Play Tennis.

Таблица 1.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

На этом этапе формируется корневой узел дерева. Подсчитывается InformationGain для каждого атрибута – кандидата и отбирается тот атрибут, у которого InformationGain наибольший в соответствии с выражениями (1) и (2).

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (1)$$

$$Entropy(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_- \qquad (2)$$

где p_+ - количество позитивных примеров, p_- - количество негативных примеров, S – набор атрибутов, $S_v = \{s \in S \mid A(S) = v\}$.

В таблице представлена выборка, которая состоит из 14 картежей. Для определения корневого узла необходимо вычислить энтропию всех исследуемых параметров – атрибутов отношения. Среди всех атрибутов информационный порог наибольший у атрибута Outlook, он определен на множестве значений: *Sunny, Overcast* и *Rain*. В соответствии с значениями параметров (1) и (2) формируется лист дерева (рис.1.)

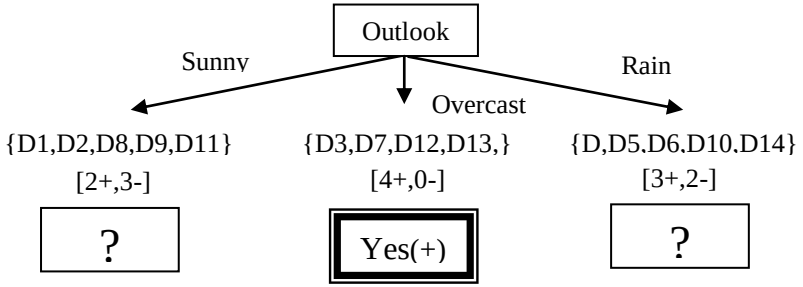

Рис.2 – Пример построения дерева решений

Аналогичные расчеты производим до тех пор, пока не будут сформированы все листья дерева. В результате получим искомое дерево решений, представленное на рис.2.

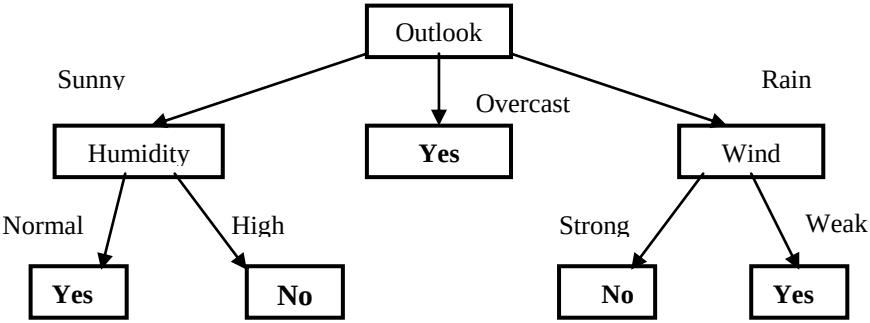

Рис.2 – Дерево решений

Рассмотренный выше метод извлечения и представления знаний обладает рядом преимуществ, среди которых можно выделить основные:

– высокая скорость процесса обнаружения знаний;

– генерация правил в для предметных областей, в которых не удается получить формальную модель представления знаний другими методами;

– интуитивно понятная классификационная модель предметной области.

Полученные результаты могут быть в дальнейшем интерпретированы в виде одной из классических моделей представления знаний – семантической сети. Семантическая сеть – это ориентированный граф, вершины которого – понятия, а дуги – отношения между ними. В качестве понятий обычно выступают абстрактные или конкретные объекты, а отношения – связи типа: «род-вид», «имеет частью», «принадлежит» и т.п. Характерной особенностью семантических сетей является обязательное наличие трёх типов отношений: «класс – элемент класса», «свойство – значение», «пример – элемент класса».

Семантическая сеть обеспечивает следующие основные функции:

- хранение сведений об объектах и связях между ними;
- поиск объектов по различным характеристикам;
- пополнение и корректировка знаний системы во время ее обучения;
- реализация различных процедур обобщения и конкретизации знаний.

В общем случае под семантической сетью понимается выражение следующего вида $S =< O, R >$, где: $R = \{R_j, j = \overline{1,k}\}$; $O = \{O_i, i = \overline{1,n}\}$; $O_i, i = \overline{1,n}$ - множество объектов конкретной предметной области; $R_j, j = \overline{1,k}$ - множество отношений между объектами; j - тип отношений.

Представим дерево решений, полученное в результате применения алгоритма ID3, в виде семантической сети. Корень дерева Outlook связан отношениями Sunny, overcast, rain с дочерними узлами. Проанализировав все компоненты дерева решений - связи и отношения, можно представить искомое множество объектов предметной области в виде:

Outlook(Sunny, Overcast, Rain)
Humidity(Normal, High)
Wind(Strong, Weak)

Таким образом, например, проблемную ситуацию «как погода влияет на игру в теннис» можно описать в виде семантической сети, структура которой представлена на рис.3.

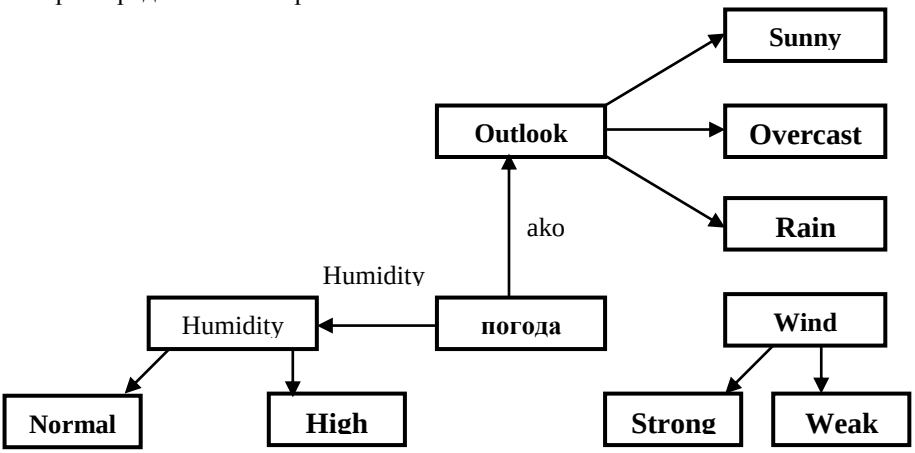

Рис.3 – Семантическая сеть

Семантической сетью можно представить формальную процедуру вывода решения, предварительно записав элементы сети в виде предикатов. Исходя из анализа дерева решений, полученное по алгоритму ID3, можно сделать вывод: на это решение влияет 3 параметра значения погоды: Outlook, Humidity, Wind. Таким образом, предикат в общем виде будет иметь вид: Идем играть (Да\нет) (Outlook, Humidity, Wind).

Следовательно:
В случае, когда идем играть
Да(Outlook=Sunny, Humidity = Normal)
Да(Outlook=Rain, Wind=Weak)
Да(Outlook=Overcast)
В случае, когда не идем играть
Нет (Outlook=Sunny, Humidity=High)
Нет (Outlook=Rain, Wind=Strong)

На основе сформулированных выше предикатов может быть представлена семантическая сеть, вид которой приведен на рис.4

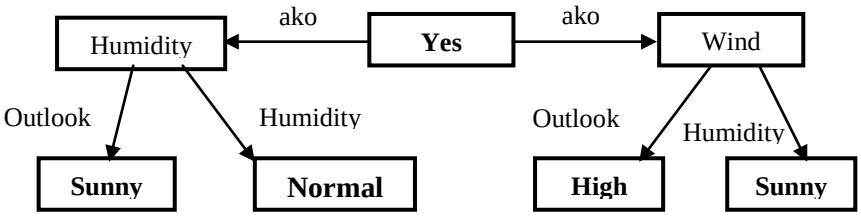

Рис.4 – Построение семантической сети

Следует заметить, что предложенный в данной статье метод интеллектуального анализа данных и представление знаний в виде семантической сети может быть применен для крупномасштабных баз данных, схема которых может содержать более 1000 атрибутов.

Выводы

В рассмотренной статье предложен комбинированный метод извлечения знаний из данных, в том числе из сильно нормализованных – реляционных баз данных. Предлагаемый подход основан на применении одного из методов анализа данных - построение дерева решений и представление полученных знаний в виде семантической сети. Решение поставленной задачи обнаружения знаний в данных, позволит повысить качество принимаемых управленческих решений.

ЛИТЕРАТУРА

1. Касаткіна, Н. В. Інформаційні системи та їх застосування [Текст] : монографія / Н. В. Касаткіна, Л. А. Пономаренко, В. О. Філатов. – К. : ПП «Аверс», 2008. – 142 с.
2. Касаткина, Н. В. Методы хранения и обработки нечетких данных в среде реляционных систем [Текст] / Н. В. Касаткина, С. С. Танянский, В. А.

Филатов // Автоматика. Автоматизація. Електротехнічні комплекси та системи. – Херсон : ХНТУ, 2009. – вип. 2 (24) . – С. 84–90.
3. Методы и модели анализа данных: OLAP и Data Mining [Текст] / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. СПб. : БХВ-Петербург, 2004. − 336 с.

Пономаренко Леонид Анатольевич
Доктор технических наук, профессор, Киевский национальный экономический университет им. В. Гетьмана
Филатов Валентин Александрович
Доктор технических наук, профессор кафедры Искусственного интеллекта Харьковского национального университета радиоэлектроники
Шевякова Алина Юрьевна
Студентка факультета компьютерных наук Харьковского национального университета радиоэлектроники

Л.А. Пономаренко, В.А. Філатов, А.Ю. Шевякова

ПРО ОДИН ПІДХІД ДО ВИРІШЕННЯ ЗАДАЧІ ВИДОБУВАННЯ ЗНАНЬ З ДАНИХ

Стаття присвячена дослідженню методів інтелектуального аналізу даних. Запропоновано комбінованих підхід видобування знань з даних, зокрема сильно нормалізованих – реляційних баз даних. Підхід побудовано на використанні методу аналізу даних – побудови дерева рішень та відображення таким чином видобутих знань за допомогою семантичних мереж. Вирішення поставленої задачі видобування знань з даних дозволяє підвищити рівень рішень що приймаються.

Л.А. Пономаренко, В.А. Филатов, А.Ю. Шевякова

ОБ ОДНОМ ПОДХОДЕ К РЕШЕНИЮ ЗАДАЧИ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ДАННЫХ

Рассматриваемая статья посвящена исследованию методов интеллектуального анализа данных. Предложен комбинированный метод извлечения знаний из данных, в том числе из сильно нормализованных – реляционных баз данных. Подход основан на применении метода анализа данных - построение дерева решений и представление полученных знаний в виде семантической сети. Решение поставленной задачи обнаружения знаний

в данных, позволит повысить качество принимаемых управленческих решений.

L.A. Ponomarenko, V.A. Filatov, A.Y. Shevjakova

ON ONE APPROACH TO SOLVING PROBLEMS OF KNOWLEDGE EXTRACTION FROM DATA

This article is devoted to research methods of data mining. In the given work proposed combined method for extracting knowledge from data, including the highly normalized – relational databases. The approach is based on the method of data analysis – building a decision tree and presentation of knowledge in the form of semantic network. The solution of the problem of knowledge discovery in data, will improve the quality of management decisions.