Krassimir Markov, Vitalii Velychko, Oleksy Voloshin
(editors)

Natural and Artificial Intelligence

ITHEA
SOFIA
2010

Krassimir Markov, Vitalii Velychko, Oleksy Voloshin (ed.)

Natural and Artificial Intelligence

ITHEA®

Sofia, Bulgaria, 2010

ISBN 978-954-16-0043-9

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book is engraved in prof. Zinoviy Lvovich Rabinovich memory. He was a great Ukrainian scientist, co-founder of ITHEA International Scientific Society (ITHEA ISS). To do homage to the remarkable world-known scientific leader and teacher this book is published in Russian language and is concerned to some of the main areas of interest of Prof. Rabinovich.

The book is opened by the last paper of Prof. Rabinovich specially written for ITHEA ISS. Further the book maintains articles on actual problems of natural and artificial intelligence, information interaction and corresponded intelligent technologies, expert systems, robotics, classification, business intelligence; etc. In more details, the papers are concerned in:conceptual problems of the natural and artificial intelligent systems: structures and functions of the human memory, ontological models of knowledge representation, knowledge extraction from the natural language texts;network technologies;evolution and perspectives of development of the mechatronics and robotics;visual communication by gestures and movements, psihology of vision and information technologies of computer vision, image processing;object classification using qualitative characteristics;methods for comparing of alternatives and their ranging in the procedures of expert knowledge processing;ecology of programming – a new trend in the software engineering;decision support systems for economics and banking;systems for automated support of disaster risk management;and etc.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0043-9

C\o Jusautor, Sofia, 2010

МОДЕЛИРОВАНИЕ И АНАЛИЗ ФАКТОВ И СВЯЗЕЙ МЕЖДУ НИМИ

Павел Мальцев

***Аннотация**: В данной работе предложен подход к моделированию и анализу фактов, записи о которых имеются в хранилище данных, и связей между ними. Описаны основные положения формальной математической теории, лежащей в основе предлагаемого подхода. Программные компоненты, реализующие предложенный подход, будут включены в программный комплекс BiP, предназначенный для многомерного анализа данных, получаемых из гетерогенных источников, и позволяет упростить разработку приложений Business Intelligence. Кроме того, планируется использовать описанный подход при разработке исследовательского портала "Инновационное развитие регионов".*

***Keywords**: исчисление фактов, Business Intelligence, BI, бизнес-анализ, Data Mining, Reporting, системы поддержки принятия решений, DSS, информационно-аналитические системы.*

***ACM Classification Keywords**: H.4 Information Systems Applications: H.4.2 Types of Systems – Decision support (e.g., MIS).*

Введение

В настоящее время ведётся разработка исследовательского портала, основной целью создания которого является интеграция информации об инновационной активности регионов, создание и апробация различных моделей инновационного развития, организация коллективной работы исследователей.

Проект создания портала "Инновационное развитие регионов" ставит задачу разработки информационно-аналитической системы, реализующей сбор, хранение, представление и анализ данных об инновационной активности регионов. При разработке портала особенно важно иметь инструмент, позволяющий извлечь максимум новых знаний из всего богатства данных, накопленных в хранилище портала. В настоящее время при решении подобных задач широко применяются методы интеллектуального анализа данных.

Пользователь портала, решая поставленные задачи, может задать вопрос: "Как связанны между собой факты, записи о которых имеются в хранилище данных?". Современные методы интеллектуального анализа данных позволяют определить лишь характер ранее выявленной связи, таким образом, само наличие некой связи между событиями должно быть установлено заранее. Например: при поиске ассоциаций, следует явно указать, между какими событиями требуется искать ассоциации. Гораздо полезнее и удобнее для пользователя иметь средство, которое позволяло бы *обнаруживать новые связи* или *проверять гипотезы исследователя о наличии связей между фактами*.

В работе представлен подход к моделированию и анализу фактов, записи о которых имеются в хранилище данных, и связей между ними. Этот подход планируется реализовать в программном комплексе BiP. Данный программный комплекс используется в качестве платформы при разработке портала «Инновационное развитие регионов» (информацию о комплексе BiP можно получить в [1]). В статье детально описаны основы формальной математической теории лежащей в основе предлагаемого подхода.

Общие сведения о предлагаемом подходе

Часто мы сталкиваемся с задачей исследования динамики развития той или иной системы. Пусть мы можем фиксировать отдельные статические состояния исследуемой нами системы. Чем характеризуются данные статические состояния? Очевидно, что они характеризуются какими-либо постоянными

значениями параметров исследуемой системы. Что же заставляет систему менять своё состояние? Очевидно, что система меняет своё состояние вследствие возникновения неких событий (внутренних или внешних по отношению системе). Следует понимать, что возникновение того или иного события определяет дальнейшее развитие всей системы.

Знания о том, как те или иные события влияют на развитие системы могут помочь в решении задачи прогнозирования дальнейшего развития системы. Кроме того, обладая данными знаниями можно прогнозировать возникновение событий в будущем.

В каждый момент времени в любой системе может происходить огромное количество событий, но не все события «полезны» для анализа. Поэтому введём понятие факта как существенного с точки зрения исследователя события. *Фактом* называется *значимое для анализа событие*.

В современных информационных системах накоплено великое множество записей о свершившихся события – фактах. Особый интерес представляют не столько сами факты, сколько связи между ними. Но, к сожалению, информация о связях между фактами в явном виде в информационных системах содержится редко. Умение проводить качественный анализ связей между фактами позволит извлечь новые, скрытые знания о причинно-следственных связях, что позволит решать задачу прогнозирования на более высоком уровне.

Выделим три типа связей между фактами: структурные, семантические, неявные. Рассмотрим характерные черты каждого типа связей:

1. *Структурные связи*. Информация о структурных связях заложена в структуре самой базы данных (БД), т.е. они явно выделены. Приведём пример явной связи. Пусть в БД содержатся данные о величине государственного финансирования тех или иных исследований и величине привлеченных инвестиций на выполнение данных исследований. Связь между фактами изменения величины государственного финансирования и размеров частных инвестиций является примером явной связи, т.к. заложена в структуру самой БД. Структурные связи не вызывают затруднения при анализе, т.к. они всегда очевидны. Более того, информационные системы всегда разрабатываются с учётом этих связей. Очевидность данных связей делает их менее ценными для решения поставленной задачи.

2. *Семантические связи*. В отличие от структурных эти связи менее явны, т.к. заложены в сами данные, в их семантику. Например: связь между ростом размера государственного финансирования исследований и ростом средней заработной платы исследователей явно в базе данных не выделена, но данная связь известна, очевидна и определяется самими данными.

3. *Неявные связи*. Знания о неявных связях являются самыми ценными для исследователей. Отличительной чертой неявных связей является то, что они неизвестны и ни в каком виде не выделены в данных, содержащихся в БД.

Поиск неявных связей и является целью предлагаемого подхода. Суть описываемого подхода – в *автоматическом построении математической модели фактов отдельной предметной области*, на основе накопленной статистики. Важную роль здесь играют так называемые *статистические шаблоны*. Данные шаблоны позволяют идентифицировать основные структуры, такие как следствия, обобщения, совокупности и т.д., то есть *структуры модели фактов*. Важной особенностью является то, что база статистических шаблонов *может редактироваться*.

Исследователь здесь обладает широким полем для эксперимента. Для описания моделей фактов предлагается формальная математическая теория – «*исчисление фактов*». По своей сути модель фактов представляет *онтологию*. Данная модель может быть впоследствии дополнена или отредактирована пользователем (экспертом), а внесённые пользователем изменения могут быть проверены на имеющихся

данных. Таким образом, описанный подход позволяет не только *автоматически стоить онтологию фактов*, но и *осуществить поддержку эксперта в проверке его гипотез*.

Описание подхода следует начать с определения понятия «исчисление фактов», которое служит основой представленного подхода.

Исчисление фактов

Каждый факт характеризуется набором атрибутов, комбинация значений которых однозначно идентифицирует факт среди остальных. Для разных фактов наборы их атрибутов могут отличаться, но каждый факт обязательно обладает атрибутами пространства и времени.

Атрибут пространства определяет точку (область) пространства в которой произошло событие (факт).

Атрибут времени определяет момент времени, в который произошло событие (факт).

Основой для проводимого анализа являются данные из хранилища данных. Как известно, сущностным свойством хранилищ данных является *поддержка хронологии*. Таким образом, исходные данные уже содержат атрибут времени, чего нельзя сказать об атрибуте пространства. Если данные не содержат атрибута пространства, то будем считать, что все факты происходят в одной точке пространства.

Будем называть *структурой факта* набор атрибутов данного факта. Для каждого атрибута должно быть определено множество допустимых значений. Данное множество и будет характеризовать атрибут. Будем обозначать атрибуты строчными буквами латинского алфавита, при необходимости будем использовать индексы.

Атрибут времени будем обозначать t, а атрибут пространства – g.

Множество допустимых значений атрибута x будем обозначать $D(x)$. Если $D(x)$ конечно, то будем говорить, что x имеет *категориальный тип*. Будем считать, что атрибуты x_1 и x_2 одинаковы, если $D(x_1) = D(x_2)$, то есть множество допустимых значений должно отражать смысл атрибута, и здесь в первую очередь важен состав множества, а не его мощность.

Таким образом, факт представляется совокупностью атрибутов с зафиксированными значениями.

Факты будем обозначать строчными буквами греческого алфавита: $\alpha, \beta, \gamma, \ldots$. Для обозначения структуры факта, т.е. набора его атрибутов, будем использовать следующее обозначение:

$$\alpha = (t, g, x_1, x_2, ..., x_n),$$

где $t, g, x_1, x_2, ..., x_n$ – атрибуты факта α. Таким образом, *фактом будем называть совокупность* n *атрибутов*. Будем полагать, что n – число из множества натуральных чисел.

Будем говорить, что наборы атрибутов $(x_1, x_2, ..., x_n)$ и $(y_1, y_2, ..., y_m)$ подобны, если выполняются следующие условия:

1) $n = m$;

2) $(\forall x_i \mid i \in \overline{1,n})(\exists j \in \overline{1,m}) : D(x_i) = D(y_j)$.

Введём следующие обозначения:

1) G – множество всех фактов;
2) $\alpha \; like \; \beta$ (будем читать, как "α подобен β") – *отношение подобия фактов*: будем говорить, что два факта подобны, если их наборы атрибутов подобны.

Следует заметить, что отношение подобия обладают свойствами транзитивности и коммутативности.

Далее приведём основное свойство структуры факта:

$$(\forall \alpha = (x_1, x_2, ..., x_n) \mid \alpha \in G, n \in N)(\forall i, j \in \overline{1,n} \mid i \neq j) : D(x_i) \neq D(x_j).$$

*Свойство транзитивности отношения подобия.*Если $\alpha like \beta$ и $\beta like \gamma$, то $\alpha like \gamma$.

Доказательство:

Пусть: $\alpha = (x_1, x_2, ..., x_n), \quad \beta = (y_1, y_2, ..., y_m), \quad \gamma = (z_1, z_2, ..., z_k)$.

$\alpha like \beta \Rightarrow n = m$, $\beta like \gamma \Rightarrow m = k$, Таким образом $n = k$ (*).

$\alpha like \beta \Rightarrow (\forall x_i \mid i \in \overline{1,n})(\exists j \in \overline{1,m}) : D(x_i) = D(y_j)$ (**)

$\beta like \gamma \Rightarrow (\forall y_j \mid j \in \overline{1,m})(\exists p \in \overline{1,k}) : D(y_j) = D(z_p) \overset{(**)}{\Rightarrow}$

$\overset{(**)}{\Rightarrow} (\forall x_i \mid i \in \overline{1,n})(\exists p \in \overline{1,k}) : D(x_i) = D(z_p) \overset{(*)}{\Rightarrow} \alpha like \beta$ ■

Свойство коммутативности отношения подобия. Если $\alpha like \beta$, то $\beta like \alpha$.

Доказательство:

Пусть $\alpha = (x_1, x_2, ..., x_n), \quad \beta = (y_1, y_2, ..., y_m)$.

Предположим обратное:

- $\overline{\beta like \alpha} \Rightarrow (\exists y_j \mid j \in \overline{1,m})(\forall i \in \overline{1,n}) : D(x_i) \neq D(y_j) \overset{\alpha like \beta}{\Rightarrow}$

  $\overset{\alpha like \beta}{\Rightarrow} (\exists i, j \in \overline{1,n})(\exists y_p \mid p \in \overline{1,m}) : D(x_i) = D(y_p) = D(x_j) \Rightarrow (\exists i, j \in \overline{1,n}) : D(x_i) = D(x_j)$

Но это противоречит основному свойству структуры факта. Таким образом, мы пришли к противоречию, а значит, наше предположение было неверным. Тем самым мы доказали свойство коммутативности отношения подобия. ■

Введём понятие класса фактов.

Множество фактов A будем называть *классом фактов*, если для него верно

$$(\forall \alpha, \beta \in \mathrm{A} \mid \alpha \neq \beta)(\forall \gamma \in G \setminus \mathrm{A}) : (\alpha like \beta) \& \overline{(\alpha like \gamma)}.$$

Таким образом, *класс фактов – это множество подобных фактов*, такое, что любой факт, не принадлежащий классу, не подобен ни одному факту из данного класса.

Классы фактов будем обозначать заглавными буквами греческого алфавита. Следует заметить, что каждому кубу в хранилище данных соответствует определённый класс фактов [2].

К сожалению, мы не всегда можем точно сказать, свершился ли факт к некоторому моменту времени t, тем более, если данный момент ещё не настал. Чаще всего, можно лишь *оценить вероятность* того, что тот или иной факт свершился (свершится) к установленному моменту. Вероятность свершения факта α к некоторому моменту времени t будем обозначать следующим образом: $P_t(\alpha), \alpha \in G$.

Предметом наших исследований являются не отдельные факты, а *причинно-следственные связи между ними*, поэтому нам важно будет знать, какова вероятность свершения некоторого факта β при условии

свершения факта α. Такую *условную вероятность* свершения факта будем обозначать следующим образом: $P_\alpha(\beta), \alpha, \beta \in G$.

Теперь можно перейти к рассмотрению *отношения следствия*. В основе предлагаемого подхода лежит предположение о том, что все факты связаны причинно-следственными связями. Другими словами: у любого факта α_1 есть причина – факт α_0. Положим также, что существует один и только один факт, не имеющий причин, будем обозначать его o. Отношение следствия будем обозначать следующим образом: $\alpha \to \beta$, где α – причина, а β – следствие.

Введём следующие аксиомы отношения следствия:

$$1^\circ \ (\forall \alpha \in G \mid \alpha \neq o) : o \to \alpha$$

$$2^\circ \ \alpha \to \beta \Rightarrow \overline{\beta \to \alpha} \quad \alpha, \beta \in G$$

$$3^\circ \ \alpha \to \beta, \ \beta \to \gamma \quad \alpha, \beta, \gamma \in G$$

$$4^\circ \ (t', x_1, ..., x_n) \to (t'', y_1, ..., y_m) \Rightarrow t' \leq t''$$

Следует сделать замечание относительно последней аксиомы. Мы не будем описывать множество допустимых значений атрибута времени и, тем более, не будем определять отношение “меньше либо равно” на данном множестве. Важно здесь понимать, что факт-следствие всегда свершается не ранее факта-причины.

Исследователь при проведении анализа не самой сложной системы может столкнуться с огромным количеством фактов. Анализ связей между отдельными фактами не поможет извлечь из данных каких-либо ценных знаний. Исследователю важно иметь *механизмы обобщения* и *группировки фактов*.

Группировкой фактов будем называть способ объединения фактов под символом одного факта.

Первым примером группировки является *совокупность* фактов. Представим ситуацию, когда мы можем представить факт в виде группы более простых фактов. Например: факт α подразумевает наступление фактов: $\alpha_1, \alpha_2, ..., \alpha_n$. Будем говорить, что α является совокупностью фактов $\alpha_1, \alpha_2, ..., \alpha_n$.

Совокупность будем обозначать как *произведение*:

$$\alpha = \alpha_1 \times \alpha_2 \times ... \times \alpha_n \text{ или } \alpha = \prod_{i=1}^{n} \alpha_i .$$

Факты, представимые в виде совокупности, будем называть *составными*, а факты, которые нельзя представить в виде совокупности, будем называть *атомарными* фактами.

Теперь представим себе ситуацию, когда под знаком некоторого факта β будем понимать факт свершения хотябы одного факта из некоторой группы $\beta_1, \beta_2, ..., \beta_n$.

Такую группировку будем называть *суммой фактов* и обозначать следующим образом:

$$\beta = \beta_1 + \beta_2 + ... + \beta_n \text{ или } \beta = \sum_{i=1}^{n} \beta_i .$$

Группу фактов будем называть *независимыми* друг от друга, если любая пара фактов из данной группы не связанна отношением следствия.

Для *независимой группы* фактов $\alpha_1, \alpha_2, ..., \alpha_n$ справедливы следующие *свойства*:

$$1^{\circ}\ P(\prod_{i=0}^{n}\alpha_i) = \prod_{i=0}^{n} P(\alpha_i)$$

$$2^{\circ}\ P(\sum_{i=0}^{n}\alpha_i) = \sum_{i=0}^{n} P(\alpha_i)$$

Будем называть *обобщением* способ объединения некоторой группы фактов под знаком одного факта таким образом, что факт, построенный в результате обобщения, обладает общими для группы свойствами.

- Пусть факт α является результатом обобщения группы фактов $\alpha_1, \alpha_2, ..., \alpha_n$, данное утверждение будем записывать следующим образом:

$$\{\alpha_1, \alpha_2, ..., \alpha_n\}\ is\ \alpha \text{ или } \alpha_i\ is\ \alpha \quad i = \overline{1,n}.$$

Для обобщения справедливы следующие *основные свойства обобщения*:

$$1^{\circ}\ (x_1 = c_1, x_2 = c_2, ..., x_m = c_m, x_{m+1}, ..., x_n)\ is\ (x_1 = c_1, x_2 = c_2, ..., x_m = c_m)$$

$$2^{\circ}\ (\forall \alpha \mid \alpha \to \beta; \gamma\ is\ \beta) : \alpha \to \gamma$$

$$3^{\circ}\ (\forall \beta \mid \alpha \to \beta; \gamma\ is\ \alpha) : \gamma \to \beta$$

Обобщение является мощным инструментом анализа, которое позволяет отсеч второстепенные свойства фактов и заглянуть глубоку в их суть.

Заключение

В данной работе рассмотрены базовые положения формальной математической теории, которая лежит в основе подхода к моделированию и анализу фактов и причинно-следственных связей между фактами. Описанный подход – основа для разработки программных средств, которые обеспечивают исследователей возможностями обнаруживать связи между фактами или проверять гипотезы о наличии связей. Программные средства проходят апробацию при создании исследовательского портала «Инновационное развитие регионов».

Благодарности

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Библиография

[1] Мальцев П. Моделированиемногомерныхданныхвсистеме METAS BI-PLATFORM // Advanced Studies in Software and Knowledge Engineering: International Book Series / Sofia, 2008. С. 173-180.

[2] Мальцев П.А., Лядова Л.Н. Формализация многомерной модели данных // Математика программных систем: Межвузовский сб. науч. тр. / Перм. ун-т. Пермь, 2006. С. 74-87.

Сведения об авторе

Павел Мальцев *– Пермский государственный университет, аспирант кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, 15; e-mail: pavel_maltsev@mail.ru.*