КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

УДК 004. 519

Ю.В. КУК, Е.И. ЛАВРИКОВА

СИСТЕМА ПОЛУЧЕНИЯ НОВЫХ ЗНАНИЙ НА ОСНОВЕ СТРУКТУРНО-ПРЕДИКАТНОЙ МОДЕЛИ ЗНАНИЙ

Данная работа посвящена дальнейшей разработке теории систем, предназначенных для извлечения знаний из экспериментальных данных [1]. Ее цель – разработка эффективного математического метода получения новых знаний о составе сложных объектов, обладающих теми или иными свойствами. Работа ориентирована на решение важной прикладной задачи – проектирование состава соединений с нужными свойствами.

В работе [1] для получения новых знаний в форме продукционных правил, во-первых, использовалось понятие переменного предиката, который может принимать множество значений – так называемых предикатных констант – предикатов в общепринятом смысле, во-вторых, – понятие расстояния между предикатами. Оба этих понятия получили дальнейшее развитие в настоящей работе. Однако в отличие от вышеупомянутых работ в данной статье рассматриваются предикаты с предметными областями, состоящими из объектов, имеющих внутреннюю структуру, т. е. объекты из предметных областей предикатов предполагаются сложными (составными), в то время как раньше они считались цельными. Составные части сложного объекта будем называть первичными объектами [2], а предикаты, обозначающие свойства и отношения первичных объектов, будем называть первичными предикатами. О свойствах и отношениях первичных объектов, входящих в состав сложных объектов, как правило, также известна некоторая информация, которая должна быть использована в процедурах получения новых знаний о составе

J. Kuk, H. Lavrikova

NEW KNOWLEDGE ACQUISITION SYSTEM ON THE BASIS OF STRUCTURAL-PREDICATE MODEL OF KNOWLEDGE

The technique of reception of new knowledge by a conclusion by analogy in semantic networks "object -a predicate" which is based on measurement of distances between groups of predicates is offered. The developed technique can be used at designing compound objects with required properties.

Предложена методика получения новых знаний выводом по аналогии в семантических сетях «объект – предикат», которая основана на измерении расстояний между группами предикатов. Разработанная методика может быть использована при проектировании составных объектов с требуемыми свойствами.

сложных объектов, обладающих теми или иными свойствами. Предлагаемая в работе процедура получения таких знаний основана на измерении расстояний между группами свойств и отношений первичных объектов, или на языке логики, между группами предикатов при некоторой их интерпретации. Мера, введенная в работе [1] для измерения степени близости предикатов, не может быть непосредственно перенесена на группы предикатов. Поэтому в данной работе вводится специальная мера, оптимальная по критерию максимального различения разных групп предикатов, для измерения расстояний между ними.

Различные виды знаний образуют иерархическую систему, отдельные элементы которой связаны структурными и семантическими связями [2]. Поэтому систему знаний удобно представлять в виде семантической сети, вершины которой соответствуют понятиям, а дуги – отношениям между понятиями. В работе рассматривается специальный тип сетевой структуры – сеть «объект – предикат», в которой свойства объектов и их отношения описываются с помощью предикатов. Эта сеть представляет собой ориентированный ациклический граф, в котором вершины соответствуют объектам и предикатам. Рассматриваются два типов объектов: составные и первичные и два типов предикатов: предикаты составных и первичных объектов. Сеть «объект – предикат» является дальнейшим обобщением сети «объект – свойство» [2]. Обобщение состоит в том, что рассматриваются не только свойства объектов, но также и их отношения. Например, двуместный предикат «разность температур кипения двух веществ больше Δ » описывает некоторое отношение между двумя объектами.

Всюду в работе под термином предикат понимается переменный предикат, представляющий собой своего рода переменную, значениями которой являются так называемые предикатные константы – предикаты в общепринятом смысле [1]. Например, предикат «цвет» следует рассматривать как переменный предикат. Он принимает следующие значения: «красный», «оранжевый», «желтый» и т. д., и эти его значения являются унарными предикатными константами. Семантическая сеть «объект – предикат» – это четырехслойный граф пирамидальной сети, отдельные слои которого образуют его вершины. Схематически эта сеть аналогична сети «объект – свойство» [2] с той лишь разницей, что вместо атрибутов объектов рассматриваются их предикаты. Обозначим P, A, S, V следующие множества вершин этой сети. Первый слой P соответствует предикатам, обозначающим свойства и отношения первичных объектов. Элементы P – это первичные предикаты. Второй слой A соответствует наименованиям первичных объектов. Они составляют предметные области первичных предикатов при их интерпретации. Третий слой S соответствует наименованиям составных объектов, четвертый V – предикатам, обозначающим свойства и отношения составных объектов. Элементы V – это предикаты составных объектов, их предметные области – составные объекты. Дуги нижнего и верхнего ярусов соединяют вершины, представляющие объекты, с вершинами, представляющими предикаты, и направлены от первичных и составных объектов к предикатам. Они используются при интерпретации предикатов. Пусть ω обозначает крат-

ность некоторого предиката. Тогда наличие ω дуг, исходящих от ω объектов и сходящихся в данном предикате, соответствует логическому значению предиката «истина» при подстановке этих объектов в предикат, и значению «ложь» – при подстановке объекта в предикат при отсутствии дуги, соединяющий данный объект с предикатом. Дуги среднего яруса соединяют вершины, соответствующие первичным объектам, с вершинами, представляющими составной объект. Первичные элементы, от которых исходят дуги, входят в состав тех составных объектов, в котором эти дуги заканчиваются.

Знания в сети «объект – предикат» получают выводом по аналогии. Вывод рассуждений по аналогии – это вывод, основанный на перенесении рассуждений из исследованной области на гомоморфную область, т.е. область в некотором смысле похожую на исследованную. В работе в качестве исследованной области для логического вывода рассуждений по аналогии выступают две группы сложных объектов G_1 и G_2 некоторой предметной области. Группа G_1 состоит из объектов, каждый из которых обладает хотя бы одним из требуемых свойств $V+$. В группу G_2 входят соединения из предметной области, которые обладают нежелательными свойствами $V-$. Задача состоит в том, чтобы наилучшим образом построить гомоморфную область, т.е. требуется сконструировать объекты группы G_3, для которых с максимальной вероятностью можно сделать следующий логический вывод рассуждений по аналогии: объекты группы G_3 суммарно обладают свойствами объектов группы G_1 и не обладают свойствами объектов группы G_2. Такой вывод есть ни что иное, как некоторое новое знание о составе объектов, обладающих нужными свойствами. Очевидно, что достоверность такого знания требует дальнейшей проверки на практике. В работе предполагается, что степень подобия объектов гомоморфной области объектам исследованной области определяется степенью похожести или близости первичных предикатов соединений группы G_3 к первичным предикатам группы G_1 и степенью отличия или удаленностью от первичных предикатов группы G_2. Правило вывода по аналогии можно сформулировать следующим образом. Пусть P_1 и P_2 – множества первичных предикатов соединений групп G_1 и G_2. $V+$ и $V-$ – соответственно желательные и нежелательные свойства соединений. Пусть P_3 – первичные предикаты некоторого проектируемого соединения с неизвестными свойствами из группы G_3. Тогда, если расстояние между множествами первичных предикатов P_1 и P_3 $d(P_1, P_3) < r1$, где $r1$ – некоторый порог, то $P_3 \to (V+) \wedge (\neg V-)$ с некоторой достоверностью $q1$, т.е. проектируемое соединение будет обладать желательными свойствами. Если $d(P_2, P_3) < r2$, где $r2$ – некоторый порог, то $P_3 \to (V-) \wedge (\neg V+)$ с некоторой достоверностью $q2$, т.е. проектируемое соединение будет обладать нежелательными свойствами. Здесь исследованная область – G1 и G2, а гомоморфная ей – G3. Для примене-

ния этого правила нужно уметь измерять расстояние между множествами первичных предикатов.

Рассмотрим задачу построения меры для измерения степени близости групп предикатов. Эта мера должна обладать следующим естественным свойством: максимально различать разные группы предикатов. Меру, обладающую этим свойством, назовем *оптимальной*.

Приведем без доказательства математические выражения и утверждения, необходимые для построения этой меры.

Введем следующие понятия. Под *меткой* x_{ik} первичного переменного предиката p_i для составного объекта s_k понимается индекс той предикатной константы предиката p_i, которая принимает логическое значение «Истина» при подстановке в нее вместо аргументов первичных объектов, входящих в s_k и соединенных дугами с этой предикатной константой. *Вектором меток* $x_k = (x_{1k}, x_{2k}, \cdots, x_{Nk})$ первичных предикатов составного объекта s_k будем называть вектор евклидового пространства R_N, элементами которого являются метки для s_k всех первичных предикатов, входящих в семантическую сеть «объект – предикат».

Можно доказать, что в качестве искомой оптимальной меры можно взять обычное евклидовое расстояние между проекциями векторов меток соединений на некоторую прямую линию. Направление ее должно быть таким, чтобы проекции векторов меток из разных групп составных объектов на эту прямую должны быть удалены друг от друга настолько далеко насколько это возможно. Такой выбор направления прямой линии позволяет оптимальным образом различать разные группы составных объектов. Прямую линию W, на которую проектируются векторы меток первичных предикатов составных объектов, назовем *проективной прямой*.

Типичным вектором меток для группы составных объектов $G_1 = \{s_1^{(1)}, s_2^{(1)}, \cdots, s_K^{(1)}\}$ назовем вектор $h^{(1)} = (\bar{x}_1^{(1)}, \bar{x}_2^{(1)}, \cdots, \bar{x}_N^{(1)})$, координаты которого равны покомпонентным средним значениям меток первичных предикатов всех составных объектов, входящих в данную группу:

$$\bar{x}_1^{(1)} = \frac{1}{K}\sum_{\nu=1}^{K} x_{1\nu}^{(1)}, \ldots, \ \bar{x}_N^{(1)} = \frac{1}{K}\sum_{\nu=1}^{K} x_{N\nu}^{(1)} .$$

Центрированным вектором меток $\tilde{x}_k$ составного объекта s_k, принадлежащего группе составных объектов $G_1 = \{s_1^{(1)}, s_2^{(1)}, \cdots, s_K^{(1)}\}$, назовем вектор

$$\tilde{x}_k = (x_{1k} - \bar{x}_1^1, x_{2k} - \bar{x}_2^1, \cdots, x_{Nk} - \bar{x}_N^1) .$$

Проекция векторов меток первичных предикатов $x_k = (x_{1k}, x_{2k}, \cdots, x_{Nk})$ составного объекта s_k на некоторую прямую линию определяется по формуле

$\Pr_c x_1 = c_1 x_{11} + c_2 x_{21} + \cdots + c_N x_{N1}$, где $(c_1, c_2, \cdots, c_N)$ – косинусы углов, образуемых этой прямой с осями координат.

Разбросом относительно произвольной точки z проекций векторов меток первичных предикатов для группы составных объектов назовем суммарное расстояние этих проекций до точки z и обозначим $D_1(z)$.

Центром проекций группы составных объектов назовем среднее значение проекций векторов меток первичных предикатов данной группы. Пусть G_1 и G_2 – две группы составных объектов, состоящих соответственно из K и L составных объектов. Для каждого составного объекта из этих групп построим на основе семантической сети «объект – предикат» вектор меток его первичных предикатов. Получим $K+L$ векторов, которые в пространстве R_N отобразятся двумя множествами векторов – X_1 и X_2. Спроецируем эти множества на проективную прямую. Обозначим множества проекций X_1 и X_2 соответственно Z_1 и Z_2, а их центры – $\bar{z}^{(1)}$ и $\bar{z}^{(2)}$. Разброс относительно произвольной точки z проекций векторов меток первичных предикатов объединенной группы составных объектов $G = G_1 \bigcup G_2$ назовем *общим разбросом* обеих групп. Обозначим его $D(z)$. *Общим центром* объединенного множества проекций $Z = Z_1 \bigcup Z_2$ назовем величину

$$\bar{z} = \frac{1}{K+L}(z_1^{(1)} + z_2^{(1)} + \cdots + z_K^{(1)} + z_1^{(2)} + z_2^{(2)} + \cdots + z_L^{(2)}).$$

Вектор разности типичных векторов меток для групп составных объектов G_1 и G_2 обозначим $h = h^{(1)} - h^{(2)} = (\bar{x}_1^{(1)} - \bar{x}_1^{(2)}, \bar{x}_2^{(1)} - \bar{x}_2^{(2)}, \cdots, \bar{x}_N^{(1)} - \bar{x}_N^{(2)})$.

Построим квадратную матрицу $H = h^T h$, где верхний индекс T обозначает операцию транспонирования. Из матрицы H с элементами $h(\nu, \mu)$ сформируем матрицу H' с элементами $h'(\nu, \mu) = \frac{KL}{K+L} h(\nu, \mu)$. Обозначим $A^{(1)}$ и $A^{(2)}$ матрицы, столбики которых состоят из компонентов центрированных векторов меток первичных предикатов для соответствующих групп составных объектов:

$$A^{(1)} = \begin{pmatrix} x_{11}^{(1)} - \bar{x}_1^{(1)} & \cdots & x_{1K}^{(1)} - \bar{x}_K^{(1)} \\ \cdots & \cdots & \cdots \\ x_{N1}^{1} - \bar{x}_1^{1} & \cdots & x_{NK}^{(1)} - \bar{x}_K^{(1)} \end{pmatrix},$$

.

$$A^{(2)} = \begin{pmatrix} x_{11}^{(2)} - \bar{x}_1^{(2)} & \cdots & x_{1L}^{(2)} - \bar{x}_{lK}^{(2)} \\ \cdots & \cdots & \cdots \\ x_{N1}^{(2)} - \bar{x}_1^{(2)} & \cdots & x_{NL}^{(2)} - \bar{x}_L^{(2)} \end{pmatrix}.$$

Введем матрицы $B^{(1)} = A^{(1)} A^{(1)^T}$ и $B^{(2)} = A^{(2)} A^{(2)^T}$. Пусть $B = B^{(1)} + B^{(2)}$.

Приведем без доказательства основную теорему о значениях направляющих косинусов углов проективной прямой W.

Теорема. Для того чтобы проективная прямая W одновременно обеспечивала максимум расстояния между центрами проекций векторов меток обеих групп соединений G_1 и G_2 и минимум разброса проекций векторов меток этих групп относительно своих центров, необходимо и достаточно, чтобы вектор значений направляющих косинусов углов проективной прямой W являлся собственным вектором матрицы $B^{-1}H'$ для ее ненулевого собственного значения.

Общие принципы получения новых знаний о составе проектируемых соединений: строится проективная прямая W с использованием многоуровневых числовых предикатов. На этой прямой ищутся центры проекций векторов меток групп G_1 и G_2; строится семантическая сеть «объект-предикат». В соответствии с правилом выбора конструируются из первичных объектов новые соединения группы G_3. Правило выбора заключается в следующем: выбираются объекты, имеющие связи с первичными предикатами, с которыми имеют также связи первичные объекты группы составных объектов G_1, и отсутствуют связи с первичными предикатами, с которыми имеют связи первичные объекты группы составных объектов G_2, при этом учитываются возможные ограничения на структуру составных объектов; из G_3 исключаются соединения, у которых проекции z на W векторов меток первичных предикатов не удовлетворяют условию $\left|z_1^* - z_2^*\right| < r1|$, где $r1$ – некоторый порог. Эта процедура позволяет отсеять ошибочно выбранные соединения и упростить сам их выбор.

Правильность теории проверялась на данных, которые были взяты из работы [3] для задачи по проектированию новых химических соединений, обладающих электрооптическими свойствами. Решение осуществлялось вышеописанным способом. В результате для контрольной группы соединений с известными типами кристаллических структур получено 100% правильных ответов.

1. *Koval V.N., Kuk Yu.V.* Distances between predicates in by-analogy reasoning systems, “Information Theories and Applications” // International Jo. – 2003. – **10**, N 1. – P. 15–22.
2. *Гладун В.П.* Партнерство с компьютером. Человеко-машинные целеустремленные системы. – Киев: «Port-Royal», 2000. – 128 с.
3. *Величко В.Ю.* Розв’язання дослідницьких задач в дискретних середовищах методами виведення за аналогією. – Дис. на соискание ученой степени кандидата технических наук. – Киев, 2003. – 150 с.