Формалізація Семантичного Моделювання з Використанням Теорії Комбінаторної Оптимізації

Тимофієва Н.К.
відділ комплексних досліджень інформаційних технологій
МННЦІТіС НАН та МОН України
Київ, Україна
TymNad@gmail.com

Formalization of Semantic Modeling Using the Theory of Combinatorial Optimization

Tymofijeva N. K.
integrated research department of information technology
ISTCITS of NAS and MES of Ukraine
Kiev, Ukraine
TymNad@gmail.com

***Анотація*—Показано, що при семантичному моделюванні виникають задачі комбінаторної оптимізації, які відносяться до задач розбиття. Це – максимальне покриття ознаками певного об'єкта, кластеризація, класифікація, таксономія. Пошук необхідної інформації в базі даних проводиться за певними ознаками, які характеризують один або кілька подібних об'єктів.**

***Abstract*—It is shown that in the case of semantic modeling, problems of combinatorial optimization arise which concern the problem of partitioning. This is the maximum coverage of signs of a particular object, clusterization, classification, taxonomy. The search for the necessary information in the database is carried out according to certain signs that characterize one or more similar objects.**

Ключові слова — база даних семантичне моделювання; задача покриття; розпізнавання мовлення; кластеризація

Keywords— database; semantic modeling; problem of coverage; speech signal; clusterization

I. ВСТУП

База даних – це структурована сукупність взаємопов'язаних даних певної предметної області. Для швидкого знаходження в ній необхідної інформації ця база має бути відповідно структурована. Це стосується не лише інформації в комп'ютері, а й будь-якої інформації про об'єкти реального світу. Наприклад, для зручності знаходження потрібної книги в бібліотеці розроблено різні каталоги, які стрктуровані певним чином [1, 2].

Представлення інформації про предметну область пов'язано з моделюванням даних. На сьогодні існують різні моделі даних, які мають свої переваги та недоліки, і кожна з моделей має свою область застосування.

У теорії моделювання даних при проектуванні їхньої структури застосовується метод, який названо семантичним моделюванням. Він полягає в моделюванні структури даних спираючись на їхній зміст що важливо для інтелектуалізації різних систем. яке дозволяє визначати глибше сутність певного об'єкта.

Як інструмент семантичного моделювання використовуються різні варіанти діаграм сутність-зв'язок або ER-діаграми [1, 3]. Для правильного застосування ER-діаграм створюються різні математичні моделі, формулювання яких базується на основі таких математичних понять як теорія множин, теорія решіток, теорія графів. Тип сутності інтерпретується як множина, а сутність – як елемент цієї множини. Якщо проаналізувати задачі структуризації та пошуку інформації в базі даних, то можна побачити, що тут має місце покриття певними ознаками об'єктів, а також виникають задачі кластеризації або класифікації.

Оскільки при семантичному моделюванні мають місце задачі комбінаторної оптимізації, наведемо їхню математичну постановку.

II. ЗАГАЛЬНА ПОСТАНОВКА ЗАДАЧІ КОМБІНАТОРНОЇ ОПТИМІЗАЦІЇ

Задачі цього класу, як правило, задаються однією або кількома множинами, наприклад A та B, елементи яких мають будь-яку природу [4]. Назвемо ці множини *базовими*. Наявні два типи задач. В *першому* типі кожну з цих множин подамо у вигляді графа, вершинами якого є її елементи, а кожному ребру поставлено у відповідність число $c_{lt} \in R$, яке називають вагою ребра (R – множина

дійсних чисел); $l \in \{1,...,n\}$, $t \in \{1,...,\tilde{n}\}$, n – кількість елементів множини A, $\tilde{n}$ – кількість елементів множини B. Покладемо, що $n = \tilde{n}$. Між елементами цих множин існують зв'язки, числове значення яких назвемо вагами. Величини c_{lt} назвемо *вхідними* даними та задамо їх матрицями. В *другому* типі задач між елементами заданої множини зв'язків не існує, а вагами є числа $v_j \in R$, $j \in \{1,...,n\}$, яким у відповідність поставлено деякі властивості цих елементів, числові значення яких задаються скінченними послідовностями, що також є вхідними даними. Ці величини визначають значення цільової функції.

Для обох типів задач із елементів однієї або кількох із заданих множин, наприклад $a_l \in A$, $l \in \{1,...,n\}$, утворюється комбінаторна множина W – сукупність комбінаторних конфігурацій певного типу (перестановки, вибірки різних типів, розбиття тощо). На елементах w комбінаторної множини W вводиться цільова функція $F(w)$. Необхідно знайти елемент w^* множини W, для якого $F(w)$ набуває екстремального значення при виконанні заданих обмежень, тобто функціонал $F(w^*) = glob \underset{w \in W^0 \subset W}{extr} F(w)$, де $extr = \{\min, \max\}$, W^0 – підмножина, яка визначається обмеженнями задачі.

III. Комбінаторні конфігурації

Комбінаторною конфігурацією назвемо будь-яку сукупність елементів, яка утворюється з усіх або з деяких елементів заданої множини $A = \{a_1,...,a_n\}$ [5]. Позначимо її впорядкованою множиною $w^k = (w_1^k,...,w_{\eta^k}^k)$. Верхній індекс k ($k \in \{1,...,q\}$) в w^k – порядковий номер w^k в W, q – їхня кількість. Множину $A = \{a_1,...,a_n\}$ назвемо базовою. Під символом $w_l^k \in A$ розуміємо як окремі елементи, так і підмножини (блоки), $\eta^k \in \{1,...,n\}$ – кількість елементів у $w^k \in W$. Залежно від умови задачі η позначатимемо без індексу або з верхнім індексом η^k. Дві нетотожні комбінаторні конфігурації w^k та w^i назвемо ізоморфними, якщо $\eta^k = \eta^i$.

IV. Задача покриття об'єктів певними ознаками

При семантичному моделюванні виникає задача максимального покриття об'єкта певними ознаками, які його характеризують. Вона відноситься до задач розбиття, аргументом цільової функції в якій є розбиття n-елементної множини на підмножини.

Ознаки розділяються на такі, які характеризують лише заданий об'єкт, за якими досить просто його визначити в базі даних. В цьому випадку задача є розв'язною. Якщо однакові ознаки описують різні об'єкти, але за допомогою диференціального аналізу можна знайти потрібний об'єкт, то така задача є частково розв'язною. Якщо одні і ті ж ознаки характеризують різні об'єкти і за ними не можна ідентифікувати пошукуваний, то виникає ситуація невизначеності. В цьому разі для розв'язання поставленої задачі необхідні додаткові умови або розробляти інші правила пошуку.

Для об'єднання спільних ознак певних об'єктів у класи необхідно розв'язати задачу кластеризації чи класифікації, які також відносяться до задач розбиття та є задачами комбінаторної оптимізації.

Розглянемо детальніше задачу покриття ознаками певного об'єкту. Змоделювавши її в рамках теорії комбінаторної оптимізації можна побачити, що аргументом цільової функції в ній є розбиття n-елементної множини на підмножини як з повтореннями так і без повторень.

Нехай задано базу даних з об'єктами різної природи Позначимо їх множиною A. Задано ознаки, які хадактеризують ці об'єкти. Позначимо їх множиною B. В цій задачі необхідно вибраними ознаками, які характеризують об'єкти, оптимально їх покрити, тобто з озак утворюються кластери. В цій задачі виділимо такі підзадачі:

• об'єкти із множини A покриваються ознаками B так, щоб останні не перетиналися.

• об'єкти із множини A покриваються ознаками B так, щоб останні повністю покривали задані об'єкти. В цьому разі один і той же елемент із B може характеризувати різні об'єкти (відноситься до різних кластерів).

Для обох задач розбиття $w \in W$ утворюється з елементів скінченної множини B.

В першій задачі утворені кластери не перетинаються, тобто $w_p \bigcap w_l = \varnothing$. Задача полягає в знаходженні такого розбиття $w^* \in W$, при якому обєкт максимально покривається мінімальною кількістю ознак.

У другій задачі утворені кластери перетинаються, тобто $w_p \bigcap w_l \neq \varnothing$. Необхідно мінімізувати кількість ознак, які характеризують вибрані об'єкти так, щоб вони повністю їх покривали, а кількість однакових у різних кластерах елементів була б мінімальною.

Оговорена задача полягає в знаходженні такого розбиття $w^* \in W$, при якому об'єкт максимально покривається мінімальною кількістю ознак при виконанні умови, а саме: кількість однакових у різних кластерах елементів була б мінімальною.

В обох задачах цільову функцію необхідно оптимізувати за двома критеріями та розв'язати проблему мінімаксу (максиміну).

Пошук інформації в базі даних проводиться за певними ознаками, якими є вхідні дані, або за однією ознакою, за якою порівнянням з кількома еталонами, знаходиться один об'єкт. На прикладі розпізнавання

мовленнєвих сигналів покажемо, яким чином проводиться пошук інформації за другим варіантом.

V. Пошук еталона у базі даних, який подібний до вхідного об'єкта

Розпізнавання мови – це процес автоматичної обробки мовленнєвого сигналу з метою визначення послідовності слів, яка передається цим сигналом [5]. Цей сигнал описується послідовністю $X = (x_1, ..., x_n)$, елемент x_i якої є значення сигналу у відліку i. Довжина n різних реалізацій сигналу певного слова – різна. Для розпізнавання з реалізацій X створюється словник еталонних слів. Еталон слова словника описується послідовністю $E_h = (e_{h_1}, .., e_{q_h})$, де h – номер слова у словнику, q_h – довжина сигналу еталону слова, $h \in \{1,...,\widetilde{q}\}$, $\widetilde{q}$ – кількість еталонних сигналів у бібіліотеці.

Задача розпізнавання мовленнєвих сигналів полягає у знаходженні для сигналу X найбільш правдоподібного еталона E_h з усіх можливих еталонних сигналів. Як видно з математичної моделі, наведеній у [5], задача розпізнавання мовленнєвих сигналів досить природно розділяється на дві підзадачі: перебір еталонних сигналів і порівняння еталонного та вхідного сигналів. Оскільки тут має місце перебір варіантів, то вона відноситься до задач комбінаторної оптимізації.

Нижче побудуємо математичну модель задачі розпізнавання як задачу комбінаторної оптимізації і визначимо комбінаторну конфігурацію, яка є аргументом цільової функції.

Розглянемо задачу порівняння еталонного і вхідного сигналів. Уведемо дві базові множини $A = \{a_1, ..., a_n\}$ і $B = \{b_1,...,b_{\widetilde{n}}\}$, де $a_i = x_i \in X$, $i = \overline{1, n}$, а $b_l = e_{h_l} \in E_h$, $l \in \{1,...,\widetilde{n}\}$, $\widetilde{n} = q_h$. Вхідні дані, якими є ваги між елементами $a_i \in A$ і $b_l \in B$ задамо несиметричною матрицею $C = \left\| c_{il} \right\|_{n x \widetilde{n}}$, номера стовпців якої збігаються з нумерацією елементів $a_i \in A$, а номера рядків – з нумерацією елементів $b_l \in B$. Як описано в [5], при поелементному розпізнаванні мовленнєвого сигналу для елемента $x_i \in X$ знаходиться йому подібний $e_{h_l} \in E_h$. Оскільки з кожної базової множини A і B вибираються по одному елементу в строгому порядку, то отримана комбінаторна конфігурація є розміщення без повторень. Позначимо її $\mu^k \in M$, де M – їхня всіляка множина. Для визначення елементів $a_i \in A$ та $b_l \in B$, що вибираються з базових множин на k-му варіанті розв'язку задачі, уведемо комбінаторну (0,1)-матрицю $Q(\mu^k) = \left\| g_{il}^k(\mu^k) \right\|_{n x \widetilde{n}}$. Якщо $g_{il}^k(\mu^k) = 1$, то з множин A і B вибрана пара (a_i, b_l), в іншому разі – значення $g_{il}^k(\mu^k) = 0$. Для запису цільової функції в явному вигляді змоделюємо вхідні дані функціями натурального аргументу. Елементи матриці C подамо числовою функцією $\varphi(j)|_1^{n^*}$, а матриці $Q(\mu^k)$ – комбінаторною $\beta(f(j), \mu^k)|_1^{n^*}$, де $n^* = n \cdot \widetilde{n}$. Кількість одиниць в комбінаторній функції дорівнює $q' = \min(n, \widetilde{n})$.

Задача порівняння еталонного і вхідного мовленнєвих сигналів полягає в знаходженні такого розміщення без повторень $\mu^{k*} = (\mu_1^{k*},..., \mu_{q'}^{k*})$, для якого функціонал

$$F(\mu^{k*}) = \max_{\mu^k \in M} \sum_{j=1}^{n^*} \varphi(j)\, \beta_j(f(j), \mu^k), \qquad (1)$$

де $\sum_{j=1}^{n^*} \varphi(j)\, \beta_j(f(j), \mu^k)$ – інтегральна міра подібності, а $\varphi(j) = g_j'(a_i, b_l)$ – елементарна міра подібності, яка визначає подібність між елементами еталонного і вхідного сигналів. Аргументом цільової функції задачі (1) є розміщення без повторень.

Розглянемо задачу пошуку еталонного сигналу, який подібний до вхідного.

Позначимо A та $\widetilde{B} = \{B_1,...,B_{\widetilde{q}}\}$, базові множини, де $A = X$, а $B_l = E_{h_l}$. В цій задачі як ваги між еталонним і вхідним сигналами виступають значення інтегральних мір подібності, одержаних за виразом (1), числове значення яких подамо матрицею C'. Номера стовпців цієї матриці збігаються з номерами еталонних сигналів, розміщених у бібліотеці. Рядок у ній один і відповідає номеру один вхідного сигналу. Оскільки при порівнянні вхідного та еталонного сигналів з базових множин A і B вибираються два елементи, то утворений об'єкт є сполучення без повторень. Позначимо його $\mu'^k \in M'$, де M' – їхня всіляка множина. Уведемо комбінаторну (0,1)–матрицю $Q(\mu'^k) = \left\| g_{1l}'^k(\mu'^k) \right\|_{1 x \widetilde{q}}$. Якщо $g_{1l}'^k(\mu'^k) = 1$, то з множин A та B вибрана пара (A, B_l), в іншому разі – значення $g_{1l}'^k(\mu'^k) = 0$. Елементи матриці C' подамо числовою функцією $\varphi'(j)|_1^{n-1}$, а матриці $Q(\mu'^k)$ – комбінаторною $\beta'(f'(j), \mu'^k)|_1^{n-1}$.

Задача пошуку еталонного сигналу, який відповідає вхідному, полягає у знаходженні такого сполучення без повторень $\mu'^{t*} = (A_1, B_l)$, для якого значення заданої цільової функції було б найбільшим, тобто

$$F(\mu'^{t*}) = \max_{\mu'^k \in M} \sum_{j=1}^{n-1} \varphi'(j)\, \beta_j'(f'(j), \mu'^t), \qquad (2)$$

де $\varphi'(j) = \sum_{j=1}^{n^*} \varphi(j)\, \beta_j(f(j), \mu^k)$.

Отже, задача розпізнавання мовленнєвих сигналів розділяється на дві підзадачі, аргументом цільової функції в одній є розміщення без повторень, а у другій – сполучення без повторень. Для визначення слова чи речення, яке описує вхідний сигнал, необхідно визначити подібний сигнал-еталон у бібліотеці еталонів. Тобто, саме цей еталон відіграє роль ознак, за якими встановлюється пошукуваний об'єкт – слово. Сигнал – це ознака і еталон. Для повного покриття об'єкта ознаками створюються кілька еталонів одного і того ж слова, тобто проводиться максимальне покриття ознаками об'єкта таким чином, щоб ці ознаки максимально його характеризували.

Як видно з постановки задачі (2), пошук еталонного сигналу, подібного до вхідного, потребує повного перебору. Для зведення цієї задачі до розв'язної проведемо структуризацію бібліотеки еталонів.

VI. Структуризація об'єктів бази даних (бібліотеки еталонів)

Упорядкуємо еталонні сигнали, що відповідають заданим словам, в алфавітному порядку за такою схемою.

1. З кожного бібліотечного сигналу виділимо сегмент постійної довжини $q^{"}$, який є початком сигналу еталонного слова так, щоб він відповідав частині першої фонеми. Множину одержаних сегментів позначимо $A = \{a_1, ..., a_n\}$, а множину слів у словнику позначимо $B = \{b_1, ..., b_n\}$. Елементу $a_l \in A$ відповідає сегмент частини першої фонеми слова, яке задається елементом b_j словника.

2. Розв'язавши задачу розбиття множини A на підмножини (кластеризацію), об'єднаємо однорідні сегменти в одну підмножину $w_s^k \subset w^k$. Підмножиною $w_s^{'k} \subset w^{'k}$ позначимо підмножину слів словника $B = \{b_1, ..., b_n\}$ з подібними початковими сегментами, яка ізоморфна $w_s^k \subset w^k$, $s \in \{1, ..., \eta^k\}$. Як і в задачі розпізнавання в цьому випадку значення функції $\varphi(j) = \sum_{j=1}^{q^{"}} g_j^{'}(\tilde{a}_{jr}, \tilde{a}_{jl})$ є інтегральною мірою подібності, а $g_j^{'}(\tilde{a}_{jr}, \tilde{a}_{jl})$ – елементарна міра подібності, яка встановлюється між сегментами $a_r, a_l \subset A$, $\tilde{a}_{jr} \in a_r$, $\tilde{a}_{jl} \in a_l$.

3. Кожній одержаній підмножині $w_s^k \subset w^k$ поставимо у відповідність еталон сегмента $a_j^{'}$, який відповідає частині першої фонеми слова, що входить до $w_s^{'k} \subset w^{'k}$. Одержану множину сегментів позначимо $A^{'} = \{a_1^{'}, ..., a_{\eta}^{'}\}$. Аналогічно можна структуризувати бібліотеку еталонних сигналів по другій, третій фонемі, використавши як еталони множину сегментів $A = \{a_1, ..., a_n\}$.

Маючи еталони сегментів $a_l^{'} \in A^{'}$, упорядкованих в алфавітному порядку, задача (1)–(2) розв'язується таким чином. При пошуку еталонного сигналу в бібліотеці вирізаємо сегмент вхідного сигналу X довжиною $q^{"}$, що відповідає частині першої фонеми. Задачу (1) розв'язуємо з використанням відомих методів, наприклад, методу динамічного програмування. При цьому порівнюється сегмент вхідного сигналу довжиною $q^{"}$ з еталонними сегментами $a_j^{'} \in A^{'}$ структуризованої бібліотеки. Якщо значення функціоналу (2) найбільше для підмножини $\rho_s^k \subset \rho^k$, то пошук вхідного слова проводиться в цій підмножині словника B по другій, третій і т. д. фонемах.

Висновки

Отже, при семантичному моделюванні мають місце задачі комбінаторної оптимізації. Покриття об'єктів певними ознаками проводяться таким чином, щоб вони повністю його покривали. При цьому ці ознаки можуть характеризувати як один об'єкт, так і декілька. Серед них можна виділити розв'язні задачі та нерозв'язні. Пошук в базі даних певного об'єкта проводиться двома способами, а саме: за певними ознаками знаходиться один або кілька подібних об'єктів, або за однією ознакою, за якою порівнянням з кількома еталонами знаходиться один об'єкт.

Література References

[1] Дейт К. Введение в системы баз данных, 8-е издание: Пер.с англ. / К Дейт. – К.; М.; СПб.: Издательский дом «Вильямс», 2005. – 1328 с:

[2] Исаченко А.Н. Модели данных и СУБД / А.Н. Исаченко, С.П. Бондаренко – Минск: БГУ, 2007. – 205 с.

[3] Сільвейструк Л.М. Формалізація моделі "сутність-зв"язок": типи сутностей, типи зв"язків та їх обмеження. автореф. ... канд. фіз-мат. наук : 01.05.03. Сільвейструк Л.М. ; КНУТШ. Київ, 2009.– 19 с

[4] Тимофієва Н.К. Теоретико-числові методи розв'язання задач комбінаторної оптимізації. Автореф. дис... докт. техн. наук / Ін-т кібернетики ім. В.М. Глушкова НАН України, Київ. – 2007. – 32 с.

[5] Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов / Т.К. Винцюк. – К.: Наукова думка, 1987. – 262 с.