Issue 9, 2021

https://doi.org/10.23939/sisn2021.09.096

УДК 004.8

АНАЛІЗ МЕТРИК ДЛЯ ІНТЕЛЕКТУАЛЬНИХ ІНФОРМАЦІЙНИХ СИСТЕМ

Віктор Григорович

Національний університет "Львівська політехніка",
viktor.grigorovich@gmail.com, 0000-0002-5828-067X

Проблема побудови метрик має вирішальне значення для розв'язання задачі кількісного оцінювання як систем об'єктів довільної природи загалом, так і відношень, що описують зв'язки між складовими вказаних систем. У сучасних інформаційних системах моделюються предметні області, які містять об'єкти та системи складної будови. Мережева модель найадекватніша для описання навколишнього світу: вона відображає об'єкти та системи об'єктів довільної природи, що взаємодіють між собою. Фактично, будь-яку систему можна описати за допомогою мережевої моделі. Потрібно окремо виділити ієрархічні моделі як різновид мережевих моделей складних систем. Ієрархічні моделі дуже поширені й використовуються у різних галузях – у біології, соціології, економіці, техніці, управлінні тощо. В кожній галузі є набір своїх ієрархічних моделей. В роботі проаналізовано метрики, придатні для оцінювання інтелектуальних інформаційних систем, зокрема – систем, які основані на онтологіях, нереляційних (ієрархічних) базах даних, ненормалізованих (вкладених) відношеннях.

Ключові слова: метрика; інтелектуальна інформаційна система; онтологія; мережева модель; ієрархічна модель.

Вступ

Для кількісного оцінювання систем першочерговою є побудова відповідної метрики, що дасть змогу визначити поняття "розмір системи" та "відстань між елементами системи". Це дасть можливість точно описувати системи та взаємозв'язки між елементами систем, перейти від їх якісних до кількісних характеристик. Для дослідження та аналізу різних систем необхідні метрики, що описують кількісні характеристики самих систем. Під час дослідження окремих елементів певної системи та відношень між цими елементами важливими будуть метрики, що описують зв'язки між окремими складовими частинами систем.

Отже, проблема побудови метрик для кількісного оцінювання мережевих та ієрархічних систем, зокрема онтологій, нереляційних (ієрархічних) баз даних, ненормалізованих (вкладених) відношень, має важливе значення. Вказані метрики дадуть змогу розв'язати задачі, пов'язані із семантичним аналізом текстів – автоматичного реферування заданого тексту, автоматичного оцінювання відповідей на відкриті тестові завдання, автоматичної побудови семантичної мережі для заданого тексту тощо.

Постановка проблеми

Опубліковано багато результатів досліджень, які стосуються побудови інтелектуальних інформаційних систем, зокрема – систем, які основані на онтологіях, нереляційних базах даних, ненормалізованих відношеннях тощо. В багатьох роботах для кількісного оцінювання таких систем та зв'язків між їх елементами вводяться метрики. Зазначені метрики використовують як графові

моделі даних (ієрархічні та мережеві), так і різновиди синтаксичних метрик та моделей, основаних на порівняннях літерних рядків. Проте всі такі моделі є певною мірою частковими і характеризуються обмеженою сферою застосування. Вказані обмеження зумовлені тим, що сьогодні відсутній адекватний сучасним потребам математичний апарат, зокрема недостатньо загальних засобів оцінювання та порівняння таких різних моделей.

У цій роботі проаналізовано метрики, пов'язані із ієрархічними та мережевими моделями, зокрема онтологіями.

Формулювання цілей статті

Метою роботи є огляд, аналіз та узагальнення наявних метрик, які використовують для кількісного оцінювання інтелектуальних інформаційних систем та зв'язків між елементами таких систем.

Аналіз останніх досліджень та публікацій

Поняття міри та метрики

Для кількісного аналізу систем необхідно визначити міру та ввести метрику.

Міра – це дійсна функція $\mu(A)$ підмножини A непустої множини Ω, така, що для всіх A, для яких визначена міра, виконуються аксіоми [1]:

1) $\mu(A) \geq 0$;
2) $\mu(\varnothing) = 0$;
3) для підмножин A_n, що попарно не перетинаються: $A_i \cap A_j = \varnothing$, $i \neq j$

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n)$$

(аксіома *зліченної адитивності*).

Міра μ називається *скінченною*, якщо $\mu(\Omega) < \infty$, і *нормованою*, якщо $\mu(\Omega) = 1$.

Крім адитивних мір, існує поняття *неадитивної* міри, наприклад, зовнішні міри, які виникли в ході розвитку теорії адитивних скінченних множин функцій. Піонером у теорії неадитивних функцій є G. Choquet [2].

Основою класичної теорії міри є адитивні множини функцій. Для фіксованої множини A із σ-алгебри Σ класична міра μ: $\Sigma \rightarrow [0, +\infty]$ для кожної множини B із Σ, такої, що $A \cap B = \varnothing$ $\mu(A \cup B) - \mu(B)$, завжди дорівнює $\mu(A)$, тобто не залежить від B. Для неадитивної міри m різниця $m(A \cup B) - m(B)$ залежить від B, що може трактуватися як вплив зв'язку A та B [3].

У системах видобування та аналітичного опрацювання даних числові міри в таблицях фактів поділяють на три категорії: *адитивні міри* можуть бути агреговані за будь-яким із вимірів таблиці фактів (наприклад, кількість проданих товарів). *Напівадитивні міри* можуть бути агреговані за деяким із вимірів, але не за всіма; в загальному випадку, баланс деякої величини – напівадитивна міра, оскільки ці значення адитивні за всіма вимірами, окрім часу (наприклад, запаси – залишок товару на складі). Нарешті, деякі міри, такі як пропорції, є повністю *неадитивними* (наприклад, відсотки) [4].

Як відомо, *метрика* – це функція ρ, яка визначає відстань у метричному просторі. Множину, для якої введено поняття відстані між елементами, називають *метричним простором* [1].

Метричний простір визначається як пара (X, ρ), де X – множина, ρ – дійсна функція (її називають *метрикою*), визначена на декартовому добутку $X \times X$, така, що:

1) $\rho(x, y) = 0 \quad \Leftrightarrow \quad x \equiv y \quad \forall\, x, y \in X$ (аксіома *тотожності*);
2) $\rho(x, y) = \rho(y, x) \quad \forall\, x, y \in X$ (аксіома *симетрії*);
3) $\rho(x, z) \leq \rho(x, y) + \rho(y, z) \quad \forall\, x, y, z \in X$ (аксіома трикутника, *нерівність трикутника*).

Із вказаних аксіом випливає властивість *невід'ємності відстані*:

$0 = \rho(x, x) \leq \rho(x, y) + \rho(y, x) = 2 \cdot \rho(x, y)$

Елементи множини *X* називаються *точками* метричного простору.

Ультраметричний простір – окремий випадок метричного простору, в якому нерівність трикутника замінена посиленою (або ультраметричною) нерівністю:

$\rho(x, z) \le \max(\rho(x, y), \rho(y, z))$

Таку метрику називають *ультраметрикою*. В ультраметричному просторі не можна отримати більшу відстань додаванням менших відстаней, тобто не виконується "принцип Архімеда".

Є багато видів метрик, які визначають різні метричні простори. Розглянемо метрики, які можуть виявитися корисними для кількісного оцінювання інтелектуальних інформаційних систем.

Метрики в інформаційних системах

Метрики, які використовуються для кількісного аналізу інтелектуальних інформаційних систем, можна поділити на такі категорії: "класичні" метрики, морфологічні метрики, ймовірнісні метрики, метрики на основі адаптивних технологій, синтаксичні та семантичні метрики для онтологій.

Класичні метричні простори

У табл. 1 наведено стислий опис класичних метрик, які використовуються для кількісного оцінювання систем різної природи та їх елементів.

Таблиця 1

Основні типи метрик та приклади метричних просторів

Позначення, назва	Опис	Метрика
Дискретний метричний простір з дискретною метрикою	множина ізольованих точок	$\rho(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases}$
Простір $\mathbf{R}^1$, числова пряма	множина дійсних чисел	$\rho(x,y)=\lvert x-y\rvert$
Простір $\mathbf{R}_1^n$, "манхеттенська метрика", "метрика міських кварталів"	множина впорядкованих груп з *n* дійсних чисел; відстань між двома векторами на "шаховій дошці", якщо можна рухатись лише під прямими кутами	$\rho_1(x,y)=\sum_{k=1}^{n}\lvert y_k-x_k\rvert$, де $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$
Евклідовий *n*-вимірний простір $\mathbf{R}^n$	множина впорядкованих груп з *n* дійсних чисел; відстань обчислюється за декартовими координатами точок за допомогою теореми Піфагора	$\rho(x,y)=\sqrt{\sum_{k=1}^{n}(y_k-x_k)^2}$, де $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$
Простір $\mathbf{R}_p^n$, відстань Мінковського	множина впорядкованих груп з *n* дійсних чисел; якщо *p* = 1 – манхеттенська відстань *p* = 2 – евклідова відстань *p* = ∞ – відстань Чебишева	$\rho(x,y)=\sqrt[p]{\sum_{k=1}^{n}(y_k-x_k)^p}$, де *p* – будь-яке фіксоване число, $p\ge 1$ $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$
Простір $\mathbf{R}_\infty^n$, метрика Чебишева, sup-метрика або метрика домінування	множина впорядкованих груп з *n* дійсних чисел;	$\rho_\infty(x,y)=\max\limits_{1\le k\le n}\lvert y_k-x_k\rvert$, де $x=(x_1,x_2,\ldots,x_n)$, $y=(y_1,y_2,\ldots,y_n)$

Продовження табл. 1

Позначення, назва	Опис	Метрика
Простір *m*	множина всіх обмежених послідовностей дійсних чисел, таких, що $\lvert x_k\rvert < \infty,\ k = 1,2,...$	$\rho(x,y) = \sup\limits_k \lvert y_k - x_k\rvert$, де $x = (x_1, x_2, ..., x_n, ...)$, $y = (y_1, y_2, ..., y_n, ...)$
Простір *c*	множина всіх збіжних послідовностей дійсних чисел $x = (\xi_1, \xi_2, ..., \xi_n, ...)$, таких, що $\exists \lim\limits_{k\to\infty} \xi_k = \xi$	$\rho(x,y) = \sup\limits_k \lvert \xi_k - \eta_k\rvert$, де $x = (\xi_1, \xi_2, ..., \xi_n, ...)$, $y = (\eta_1, \eta_2, ..., \eta_n, ...)$
Простір c_0	множина всіх збіжних до нуля послідовностей дійсних чисел $x = (\xi_1, \xi_2, ..., \xi_n, ...)$, таких, що $\exists \lim\limits_{k\to\infty} \xi_k = 0$	$\rho(x,y) = \sup\limits_k \lvert \xi_k - \eta_k\rvert$, де $x = (\xi_1, \xi_2, ..., \xi_n, ...)$, $y = (\eta_1, \eta_2, ..., \eta_n, ...)$
Простір l_2	множина послідовностей дійсних чисел, таких, що $\sum\limits_{k=1}^{\infty} x_k^2 < \infty$	$\rho(x,y) = \sqrt{\sum\limits_{k=1}^{\infty} (y_k - x_k)^2}$, де $x = (x_1, x_2, ..., x_n, ...)$, $y = (y_1, y_2, ..., y_n, ...)$
Простір l_p, $p \ge 1$	множина послідовностей дійсних чисел, таких, що $\sum\limits_{k=1}^{\infty} \lvert x_k\rvert^p < +\infty$	$\rho(x,y) = \sqrt[p]{\sum\limits_{k=1}^{\infty} \lvert y_k - x_k\rvert^p}$, де $x = (x_1, x_2, ..., x_n, ...)$, $y = (y_1, y_2, ..., y_n, ...)$
Простір l_∞	множина всіх обмежених послідовностей $x = (x_1, x_2, ..., x_n, ...)$ дійсних чисел, таких, що $\lvert x_k\rvert < c_x,\ k = 1,2,...$ де c_x – константа, своя для кожної послідовності *x*	$\rho(x,y) = \sup\limits_k \lvert y_k - x_k\rvert$, де $x = (x_1, x_2, ..., x_n, ...)$, $y = (y_1, y_2, ..., y_n, ...)$
Простір $C[a,b]$	множина всіх неперервних дійсних функцій, визначених на відрізку $[a, b]$	$\rho(x,y) = \max\limits_{a\le t\le b} \lvert y(t) - x(t)\rvert$
Простір неперервних функцій з квадратичною метрикою $C_2[a,b]$	множина всіх неперервних дійсних функцій, визначених на відрізку $[a, b]$	$\rho(x,y) = \sqrt{\int\limits_a^b (y(t) - x(t))^2\, dt}$
Простір $L_2[a,b]$	множина всіх функцій *x*(*t*), інтегрованих з квадратом на відрізку $[a, b]$, для яких $\int\limits_a^b (x(t))^2\, dt < +\infty$	$\rho(x,y) = \sqrt{\int\limits_a^b (x(t) - y(t))^2\, dt}$

Закінчення табл. 1

Позначення, назва	Опис	Метрика
Простір $L_p[a,b]$, $p \geq 1$	множина всіх функцій $x(t)$, визначених на відрізку $[a, b]$, для яких $\int_a^b \lvert x(t)\rvert^p dt < +\infty$ (в значенні Лєбега)	$\rho(x,y) = \sqrt[p]{\int_a^b \lvert x(t) - y(t)\rvert^p dt}$
Відстань Махаланобіса [5]	міра відмінності між двома випадковими векторами $\vec{x}$ та $\vec{y}$ із одного розподілу ймовірностей з матрицею коваріації S	$\rho(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$
Відстань Хеммінга [6]	кількість значень, що відрізняються між двома векторами, зазвичай використовується для порівняння двох двійкових рядків однакової довжини; міра відмінності між кодовими комбінаціями (двійковими векторами) у векторному просторі кодових послідовностей, в цьому випадку відстанню Хеммінга $\rho(x, y)$ між двома двійковими послідовностями (векторами) x і y довжини n називається кількість n_d позицій, в яких вони різні	$\rho(x,y) = n_d$

У роботі [7] проаналізовано дев’ять метрик та вказано їх переваги та недоліки у разі використання для кількісного оцінювання систем. Автор розглянув метрику Евкліда, метрику косинуса – $d(x, y) = \cos(\theta) = x \cdot y/(\|x\| \cdot \|y\|)$, метрику Хеммінга, манхеттенську метрику, метрику Чебишева, метрику Мінковського, метрику Жакара [8] – $d(x, y) = 1 - |x \cap y|/|x \cup y|$, метрику гаверсинуса – описує відстань між двома точками на сфері, метрику Соренсена–Дайса [9, 10] – $d(x, y) = 2 \cdot |x \cap y|/(|x|+|y|)$.

Зазначено недоліки: перед використанням метрики Евкліда необхідно нормалізувати дані, ця метрика придатна лише за невеликих розмірностей. Недоліком метрики косинуса є те, що не враховується величина векторів, а лише їх напрямки. Метрику Хеммінга важко використовувати, коли довжини двох векторів різні. Манхеттенська метрика працює за великих розмірностей, але є менш інтуїтивно зрозумілою, ніж евклідова, і дасть більше значення відстані, бо не використовує найкоротший шлях. Метрику Чебишева зазвичай використовують у дуже конкретних випадках, вона непридатна як універсальна метрика. Метрика Мінковського має ті самі недоліки, що і метрики Манхеттена, Евкліда або Чебишева; із параметром p іноді складно працювати, оскільки пошук його правильного значення може бути доволі неефективним в обчисленні залежно від певного випадку використання. Основним недоліком метрики Жакара є те, що на неї істотно впливає розмір даних: великі набори даних можуть суттєво впливати на відстань, оскільки це може істотно збільшити об’єднання, тоді як перетин залишається подібним. Недолік міри відстані гаверсинуса – передбачається, що точки лежать на кулі. Недолік метрики Соренсена–Дайса – те, що, як і метрика Жакара, вона також перебільшує важливість множин із незначними або відсутніми позитивними показниками; домінуватиме середній показник декількох наборів – зважується кожен елемент обернено пропорційно до розміру відповідного набору, замість однакового опрацювання наборів.

Морфологічні метрики для ієрархічних дерев

N. E. Fenton та S. L. Pfleeger [11] описали набір простих морфологічних метрик для ієрархічних дерев, основаних на характеристиках графів:

- *Розмір Size* = n, де n – кількість вершин.

- *Густина взаємодії* $R = e/n$ (відношення кількості ребер до кількості вершин). Для дерева $e = n–1$.
- *Коефіцієнт розгалуження за виходом* $Fan\_out(i)$ – це кількість дочірніх вершин i-ї вершини.

Первинними характеристиками графу є кількість вершин n та кількість ребер e. Для дерева до них додаються ще дві глобальні характеристики – висота і ширина:

- *висота* (*depth*) – кількість рівнів (кількість вершин у найдовшому шляху від кореневої вершини до листової);
- *ширина* (*width*) – максимальна кількість вершин, розміщених на якомусь одному рівні дерева. *Ширина рівня* – це кількість вершин дерева на цьому рівні, тоді *ширина дерева* – це максимальна ширина на всіх рівнях (наприклад, дерево на рис. 1).

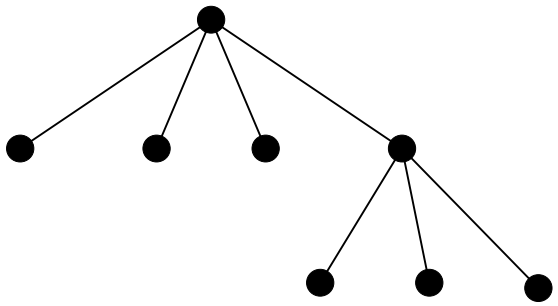

Рис. 1. Приклад ієрархічного дерева для ілюстрації морфологічних метрик: n = 8, e = 7, висота = 3, ширина = 4

Ймовірнісні (байєсові) метрики

У роботі [12] запропоновано ймовірнісний підхід, оснований на теоремі Байєса, для оцінювання ієрархічних структур під час побудови експертних систем.

Метрика для таких систем основана на теоремі Байєса: ймовірність здійснення деякої гіпотези H за наявності певних свідчень E, які підтверджують цю гіпотезу (тобто у разі настання подій E), обчислюють на основі апріорної ймовірності цієї гіпотези без свідчень-підтверджень E та ймовірності здійснення свідчень за умов, що гіпотеза правильна або хибна:

$$P(H \mid E) = \frac{P(HE)}{P(E)} \quad \Rightarrow \quad P(HE) = P(H \mid E) \cdot P(E) = P(E \mid H) \cdot P(H),$$

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)},$$

де $P(H)$ – апріорна ймовірність гіпотези H; $P(H \mid E)$ – ймовірність гіпотези H у разі настання подій E (апостеріорна ймовірність); $P(E \mid H)$ – ймовірність настання подій E, якщо істинна гіпотеза H; $P(E)$ – ймовірність настання подій E.

Г. С. Теслер у роботі [13] розвиває підхід, який запропонував К. Нейлор. Нехай $G = (V, E)$ – зв'язний граф, u та v – дві його різні вершини. Тоді відстанню між вершинами u та v буде довжина найкоротшого маршруту, яка позначається $\rho(u, v)$. При цьому виконуються всі аксіоми метрики.
Відомо, що всякий граф взаємно однозначно представляється бінарним відношенням, яке можна задати матрицею суміжності. Елементи матриці суміжності $A(G)$ мають вигляд

$$a_{ij} = \begin{cases} 1, & \text{якщо вершини з номерами } i \text{ та } j \text{ - сумісні} \\ 0, & \text{в іншому випадку} \end{cases}$$

Рангом графу G називають ранг його матриці суміжності, позначають $rank(G)$.

Якщо u – деяка вершина графу $G = (V, E)$, то величина $e(u) = \max \rho(u, v)$, $v \in V$ називається ексцентриситетом вершини u. Діаметром графу називають максимальний ексцентриситет серед його вершин і позначають $d(G) = \max e(u)$, $u \in V$.

У цьому випадку як міру можна було б використовувати діаметр графу і побудувати метрику графів на основі їх діаметрів, що еквівалентно одній із морфологічних метрик.

Дерево в теорії графів – це зв'язний ациклічний неорієнтований граф. Інший підхід визначає дерево як орієнтований граф, що містить єдиний особливий вузол (корінь дерева), решта вузлів поділяються на $n \geq 0$ множин $T_1, T_2, \ldots, T_n$, які взаємно не перетинаються; кожна з цих множин, своєю чергою, є деревом. Множини T_i, $i = 1, \ldots, n$ називають піддеревами кореня.

Як приклад ієрархії Г. С. Теслер наводить дерево хвороб людини та їх зв'язки залежно від причин їх виникнення та механізмів їх розвитку. Як основу метрики для таких систем використовують теорему Байєса. Використання такого підходу для побудови експертної системи для медичної бази знань MYCIN викладено в роботі [12]; подібний підхід для аналізу кристалічних структур хімічних сполук описано в роботі [14].

Зауважимо, що система MYCIN оперує поняттям "степінь певності": процедурні правила в ній формулюються у вигляді "**ЯКЩО … ТО … З ПЕВНІСТЮ *P***", де степінь певності – "приблизно те саме, що ми називаємо умовною ймовірністю $P(H \mid E)$ – ймовірність гіпотези H за умови, що подія E відбулася" [12]. Під час побудови системи MYCIN експерти-медики пропонували правила і вказували степінь довіри до кожного правила в діапазоні від 1 до 10 – такі експертні оцінки і стали степенем певності для відповідних процедурних правил. Отже, набір процедурних правил такої системи можна описати за допомогою орієнтованого графу, кожне ребро якого має вагу – умовну ймовірність переходу $E \rightarrow H$, тобто ймовірність гіпотези H за умови, що відбулася подія E (рис. 2):

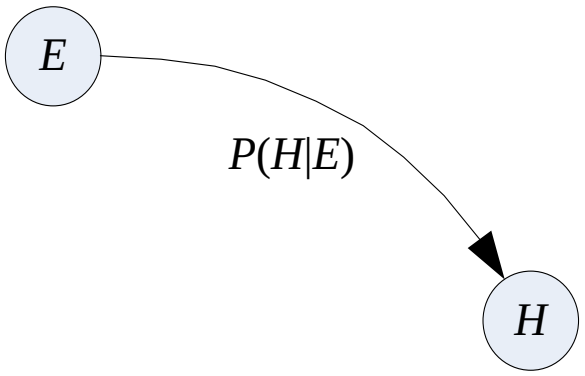

Рис. 2. Зважений орієнтований граф, для якого вага ребра – це умовна ймовірність переходу від одного вузла до іншого (ілюстрація до ймовірнісних метрик)

Очевидно, що такий підхід дає змогу оцінити лише окремі частини деякого орієнтованого графу і не придатний для порівняння різних графів між собою, бо загальна ймовірність повної системи повинна дорівнювати одиниці: $P(\ G = (V, E)\ ) = 1$.

Метрика на основі адаптивних онтологій

У роботах [15–18] запропоновано метрики на основі адаптивних онтологій для семантичних (наприклад, задач класифікації) та ознакових (наприклад, пошук релевантних прецедентів) задач.

Для семантичних задач відстань між прецедентом і ситуацією визначається як сума відстаней між "найважливішими" поняттями прецеденту та поточного випадку. Найважливіше поняття відповідає центру ваг концептуального графу, за допомогою якого подається адаптивна онтологія. Розглядають максимум три "найважливіші" поняття концептуального графу. В цьому випадку отримуємо три центри ваг i-го прецеденту pr_i^1, pr_i^2, pr_i^3 та три центри ваг поточної ситуації s^1, s^2, s^3. Відстань між прецедентом та поточною ситуацією визначається як

$$d(pr, s) = \arg\min \sum_{n=1}^{3} d_n,\ d_n = d(pr_i^j, s^k),\ j = 1,2,3,\ k = 1,2,3$$

– з дев'яти різних відстаней $d\left(pr_i^j, s^k\right)$, j=1,2,3; k=1,2,3 вибирають такі три, що їх сума була мінімальною. Отримана сума і буде відстанню між прецедентом та поточною ситуацією.

Пропоновані метрики на основі адаптивних онтологій визначають відстань у межах заданого онтологією концептуального графу.

Метрики для оцінювання онтологій

У роботі [19] запропоновано підхід машинного навчання для поєднання метрик, який використовує різні лінгвістичні та контекстні профілі для виявлення відповідності між сутностями різних онтологій. Запропоновано метрики, які стосуються термінологічних та контекстних характеристик сутностей в онтології. Автори поділяють термінологічні метрики на дві категорії: синтаксичні (string-based) та лінгвістичні (linguistic-based). Синтаксичні метрики враховують подібні символи двох літерних рядків, а лінгвістичні порівнюють значення цих рядків. Запропонований підхід полягає у таких кроках: по-перше, назви сутностей розділяються на окремі лексичні одиниці (токени), після чого токени можна порівнювати за допомогою примітивних синтаксичних та лінгвістичних метрик. По-друге, токени можуть мати різні морфологічні форми. Для цього потрібний словник-тезаурус. Автори запропонували алгоритм порівняння токенів за кількістю однакових морфологічних форм. По-третє, ім'я сутності може бути абревіатурою, акронімом чи взагалі незрозумілою послідовністю символів. Тоді сутність повинна мати зрозумілу для людини анотацію чи коментар, і в цьому випадку порівнюються ці анотації чи коментарі. Використовується модель дерева рішень для об'єднання метрик та стратегій подібності для вибору метрик та навчальних даних для процесу машинного навчання. Описано результати здійснених експериментальних досліджень, які підтверджують надійність розроблених моделей.

Робота [20] пропонує вирішення проблеми, пов'язаної зі збільшенням кількості онтологій та семантичних веб-додатків. Кількість та складність таких онтологій ускладнює розробникам вибір, які саме онтології використовувати. Для спрощення цієї проблеми пропонується алгоритм модуляризації, який може використовуватися для розділення онтологій на набори модулів. Для того, щоб оцінити якість модуляризації, запропоновано нову оцінювальну метрику, яка кількісно оцінює ефективність модуляризації онтології: однорідність модуля (MOHO) запропоновано використовувати як метрику внутрішніх характеристик понять модуля, а неоднорідність модуля (MOHE) – як оцінку взаємозалежності між модулями онтології.

Метрика MOHO використовує семантичні та структурні характеристики модульних понять, зокрема назви понять, відстані між поняттями та кількість дерев, що утворюються від модуляризації, а метрика MOHE охоплює різні аспекти відношень між модулями. Запропоновані критерії визначають зв'язність модуля, а також зв'язки між модулями. По-перше, розмір елемента модуля безпосередньо впливає на його зв'язність; автори вважають, що це залежить від окремих класів, тому зараховують лексичний рівень кожного модуля до міри однорідності модуля. По-друге, більш пов'язані концепти в межах того самого модуля показують, наскільки висока однорідність модуля, тому обчислюється кількість зв'язаних концептів, а саме кількість дерев, оскільки кожне дерево відображає один-єдиний зв'язаний компонент концептів. По-третє, глибина концептів – це важливий аспект для оцінювання однорідності модуля. Чим ближчі концепти в межах модуля один до одного, тим зв'язніший модуль. Нарешті, важлива проблема, коли одна онтологія розділяється на окремі модулі, – скільки ребер при цьому буде вирізано.

Метрика MOHO для кількісної оцінки однорідності модулів онтології містить такі окремі метрики: Semantic module homogeneity (SMH) – метрика, основана на семантичній подібності заголовка кластера (центральна частина модуля) та інших концептів модуля. Structural module homogeneity (SrMH) – міра структурної однорідності модуля, обчислюється на основі суми мінімальних шляхів між заголовком кластера та концептами модуля. Average depth of all concepts (AvgDepth) – міра середньої глибини всіх концептів, обчислюється як середнє значення довжини шляху від кожного концепту до кореня графу концептів. Number of trees in each module (NTree), – кількість дерев у кожному модулі. Ключовим питанням є те, скільки концептів пов'язані один з одним, чи вони відокремлені один від одного. Автори визначили критерій для вимірювання кількості пов'язаних концептів у кожному модулі. Якщо він низький, то демонструє більшу зв'язність модуля: відношення кількості кореневих концептів у модулі до загальної кількості концептів у цьому самому модулі.

Для визначення неоднорідності модуля (MOHE) кількісно оцінюють взаємозалежність між різними модулями. Тому розглядаються такі метрики: Relative size (RS) – відносний розмір. Бажано мати такі модулі онтології, які були б слабко пов'язані між собою, – тоді можна буде їх використовувати незалежно один від іншого. За відносним розміром забезпечують, щоб концепти онтології нормально розподілялися між модулями. Відносний розмір обчислюється як відношення суми абсолютних значень різниць розмірів різних модулів до розміру усієї онтології. DetachRel – відносна кількість відокремлених зв'язків між концептами (властивостями) із різних модулів.

У роботі [21] описано вирішення проблеми, пов'язаної із тим, що різноманітність способів концептуалізації домена призводить до створення різних онтологій із суперечливими частинами або частинами, які перекриваються. З цієї причини онтології потрібно узгодити (вирівняти). Одним з методів вирівнювання онтології є порівняння назв класів та властивостей онтологій за допомогою метрик, основаних на відстанях між літерними рядками. В літературі описано доволі багато таких метрик. Але всі вони були спочатку розроблені для різних додатків і полів, що призводить до низької продуктивності в разі застосування у новому домені. В цій роботі подано нову метрику літерних рядків для порівняння імен, яка краще працює під час вирівнювання онтологій, а також в багатьох інших задачах зіставлення полів. Подібність між двома сутностями пов'язана з їх спільністю, а також їх відмінностями. Функція спільності визначається метрикою підрядків літерного рядка. У метриці підрядків обчислюється найбільший загальний підрядок двох рядків. Цей процес продовжується із видаленням спільного підрядка та повторним пошуком наступного найбільшого підрядка, поки їх не залишиться. Сума довжин цих спільних підрядків потім масштабується з довжиною рядків. Функція різниці ґрунтується на довжині невідповідних рядків, отриманих у результаті початкового кроку узгодження. Фактично запропонована метрика літерних рядків основана на синтаксичному (а не семантичному) аналізі тексту.

Робота [22] – це огляд метрик для оцінювання якості та правильності онтологій. ONTOMETRIC, OntoQA і Protégé представляють найважливіші інструменти для оцінки онтологій. Також існують різноманітні метрики зчеплення, зв'язку та ранжування і такі методики, як OntoClean. В статті проаналізовано ці інструменти та метрики і розглянуто сучасний стан метрик онтологій.

У статті [23], яка описує деякі метрики для нормалізації онтологій, розглянуто сучасний стан та запропоновано передусім нормалізацію як попередній процес застосування структурних метрик. Цей процес нормалізації складається із п'яти етапів: найменування анонімних класів, найменування анонімних індивідів, ієрархічної класифікації та уніфікації імен, поширення індивідів до найглибших можливих класів та нормалізації властивостей об'єкта. Ця пропозиція сконцентрована на контентних метриках, основаних на платформі OntoMetric, їх запропоновано насамперед для покращення поведінки онтології або для виправлення деяких помилок.

Роботи [24] та [25] пропонують деякі метрики для ранжування онтологій. Ця пропозиція складається здебільшого із сервлета Java для опрацювання як вхідних даних певних ключових слів, які ввів користувач. Потім фреймворк здійснює пошук за допомогою методу Swoogle1 і отримує всі URI, що представляють онтології, пов'язані із цими ключовими словами. Потім фреймворк шукає у своїй внутрішній базі даних, чи ці онтології раніше були проаналізовані, та одержує їх інформацію. Нарешті, фреймворк ранжує отримані онтології.

Orme та ін. [26] запропонували набір метрик зв'язку для систем на основі онтологій, представлених в OWL. Такими метриками є: кількість зовнішніх класів (NEC), посилань на зовнішні класи (REC) та включених посилань (RI). Ця пропозиція визначає новий тип вимірювання зв'язку для розроблення системи, який визначає метрику зв'язку на основі даних онтології та її структури. Перша запропонована метрика – NEC, що представляє кількість різних зовнішніх класів, визначених поза онтологією, які використовуються для визначення нових класів та властивостей онтології. Зовнішні класи можуть містити стандартні класи, визначені як примітиви мови онтології, та класи з інших онтологій, які визначив користувач. Друга метрика REC – кількість посилань на зовнішні класи онтології. NEC – це пряма міра кількості класів онтології. REC – це пряма міра кількості

розгалужень (у цьому випадку розгалуження – різні ієрархії класів із зовнішніми коренями) в цій онтології, які є результатом зовнішніх класів. RI – це пряма міра кількості включених в онтологію посилань. Для поєднання цих метрик використовується стандартний аналізатор на основі XML.

В роботі [27] автори запропонували набір метрик з'єднання онтологій для вимірювання модульної зв'язаності онтологій OWL. Ці показники – кількість кореневих класів (NoR), кількість листових класів (NoL) та середня глибина ієрархічного дерева успадкування для всіх листових вузлів (ADIT-LN). Автори визначають NoR-метрику як загальну кількість кореневих класів, явно визначених в онтології. Кореневий клас в онтології означає, що клас не має семантичного суперкласу, явно визначеного в онтології. Метрика NoL визначається як кількість класів листів, явно визначених в онтології. Листовий клас в онтології означає, що клас не має семантичного підкласу, явно визначеного в онтології. Нарешті, ADIT-LN визначається як сума глибин усіх шляхів, поділена на загальну кількість шляхів. Глибина – це загальна кількість вузлів на шляху від кореневого вузла до листового вузла. Загальна кількість шляхів в онтології – це всі окремі шляхи від кожного кореневого вузла до кожного листового вузла, якщо існує шлях успадкування від кореневого вузла до листового вузла. Кореневий вузол – це перший рівень у кожному шляху.

Yinglong та ін. [28] запропонували інший набір метрик зв'язності онтології для вимірювання модульної зв'язаності онтологій у контексті динамічної та мінливої мережі Web. Ці метрики визначено з урахуванням принципу зв'язаності з об'єктно-орієнтованого підходу, адаптованого до онтологій. Автори зосереджуються на вимірюванні невідповідностей в онтологіях і повністю розглядають онтологічну семантику, а не структуру. Метрики, які вони пропонують, – це кількість онтологічних розділів (NOP), кількість мінімально неузгоджених підмножин (NMIS) та середнє значення аксіомних неузгодженостей (AVAI). У цій роботі також описано алгоритми для обчислення цих метрик та перевірки метрик за допомогою платформ валідації. Ці метрики орієнтовані на оцінювання якості онтологій. Автори визначають метрику NOP як кількість семантичних розділів бази знань. NMIS визначається як кількість усіх мінімально несумісних підмножин у базі знань. Цей показник корисний для вимірювання масштабу впливу неузгодженості бази знань. Третя метрика AVAI визначається як: відношення суми значень впливу неузгодженості всіх аксіом і тверджень до кардинальності (кількості елементів) бази знань. Крім того, в статті проаналізовано та підтверджено запропоновані показники. Взагалі кажучи, до переваг цих метрик належить можливість оцінювання якості узгодженості онтології.

Методологія OntoClean [29, 30] пропонує використовувати деякі визначені метавластивості: жорсткість, єдиність, ідентичність та залежність. Автори запозичили ці концепції із античних філософських аналогів. Методологія складається із присвоєння цих метавластивостей сутностям з метою надання їм логічного та смислового значення. Застосування цих метавластивостей призводить до накладення декількох обмежень на таксономічну структуру онтології та дає змогу розробити концептуальний аналіз концептів та їх обґрунтованості. І навіть більше, ця методологія дає змогу аналізувати та виявляти нелогічно узгоджені відношення.

YANG та ін. [31] запропонували метрики, які враховують еволюцію онтологій. Автори пропонують набір метрик складності, які в основному розглядають кількість, співвідношення та корелятивність понять та відношень, щоб оцінити онтології з погляду складності та її еволюції. Ці метрики поділяються на дві групи: примітивні метрики та метрики складності. Примітивні метрики включають TNOC (загальна кількість понять або класів), TNOR (загальна кількість відношень), TNOP (загальна кількість шляхів), де шлях визначається як маршрут від певного конкретного концепту до найзагальнішого концепту в онтології. Перша метрика складності визначається як середня кількість відношень на концепт, що обчислюється діленням TNOR на TNOC. Друга метрика – це середня кількість шляхів на концепт, обчислюється діленням TNOP на TNOC.

У роботі [32] розглянуто питання пошуку ефективного методу оцінювання онтологій, наведено огляд наявних методів оцінки онтології, обговорено їхні переваги та недоліки. Наведені методи оцінювання онтології можна згрупувати у чотири категорії: підходи, основані на золотому стандарті, на сукупності елементів, на завданнях та на критеріях.

Насамперед розглянуто критерії оцінювання онтологій. Оскільки онтології розглядаються як еталонні моделі, необхідно забезпечити їх оцінку з огляду на дві важливі характеристики: якість та правильність. Ці дві характеристики відповідають кільком критеріям. Точність – це критерій, який визначає правильність визначення та описів класів, властивостей та індивідуалів в онтології. Міри повноти – критерій того, наскільки предметна область належно охоплена онтологією. Стислість – критерії, які визначають, чи онтологія містить нерелевантні елементи щодо області, яку потрібно охопити. Адаптивність вимірює, наскільки онтологія передбачає своє використання. Онтологія повинна запропонувати концептуальну основу для цілої низки очікуваних завдань. Чіткість вимірює, наскільки ефективно онтологія повідомляє передбачуване значення визначених термінів. Визначення мають бути об'єктивними та незалежними від контексту. Обчислювальна ефективність вимірює здатність використовуваних інструментів працювати з онтологією, зокрема швидкість, необхідну машині висновків для виконання завдання. Узгодженість вказує, що онтологія не містить чи не допускає ніяких суперечностей.

Підходи на основі золотого стандарту, що також відомі як вирівнювання онтології або онтологічне відображення, є найбільш прямим підходом. Цей тип підходу намагається порівняти онтологію, яка досліджується, з раніше створеною еталонною онтологією, відомою як золотий стандарт. Цей золотий стандарт представляє ідеалізований результат алгоритму навчання. Однак створення відповідної золотої онтології може бути складним, оскільки вона повинна бути створена в аналогічних умовах із аналогічними цілями, що і досліджувана онтологія. З цієї причини деякі підходи створюють специфічні таксономії за допомогою експертів-людей, щоб використовувати їх як золотий стандарт. Хоча інші підходи вважають за краще використовувати надійні, популярні таксономії в подібній галузі, щоб розглядати їх як еталонні онтології, оскільки це істотно зменшує обсяг роботи.

Наприклад, Maedche і Staab (2002) розглядають онтології як двошарові системи, що складаються із лексичного та концептуального шарів. На основі такої основної моделі онтології цей підхід вимірює схожість між вивченою онтологією та онтологією сфери туризму, що моделюють експерти. Він вимірює подібність на основі поняття лексикону, опорних функцій та семантичної котопії, які детально описано в [33].

Крім того, Ponzetto і Strube [34] оцінюють похідну систематику з Вікіпедії, порівнюючи її з двома базовими таксономіями. По-перше, цей підхід відображає вивчену систематику з ResearchCyc за допомогою денотаційного картографу лексеми до концепції. Потім він обчислює семантичну схожість з WordNet, використовуючи різні сценарії та заходи: Rada et al. (1989), Ву та Палмер (1994), Лікок і Чодоров (1998) та міра Ресніка (1995).

Treeratpituk та ін. [35] оцінюють якість побудованої систематики із великого текстового корпусу, порівнюючи її з шістьма специфічними таксономіями золотого стандарту. Ці шість еталонних таксономій генеруються з Вікіпедії за допомогою запропонованого ними алгоритму GraBTax.

Zavitsanos та ін. [36] також оцінюють вивчену онтологію на основі золотого відліку. Цей новий підхід перетворює поняття онтології та їх властивості на векторне представлення простору та обчислює схожість та несхожість двох онтологій на лексичному та реляційному рівнях.

Цей тип підходу застосовують також Kashyap та Ramakrishnan [37]. Вони використовують базу даних MEDLINE як корпус документа, а тезаурус MeSH – як золотий стандарт для оцінювання побудованої систематики. Процес оцінювання порівнює сформовану таксономію з еталонною систематикою із використанням двох класів метрик: 1) метод якості вмісту: вимірює перекриття міток між двома таксономіями з метою вимірювання точності та відкликання; 2) структурний показник якості: вимірює структурну обґрунтованість етикетки (тобто коли в одній таксономії з'являються дві мітки у відносинах батько–дитина, вони повинні з'являтись у послідовному взаємозв'язку (батько–дитина або пращур–нащадок) в іншій таксономії).

Підходи на основі золотого стандарту ефективні для оцінювання точності онтології. Висока точність забезпечується правильними визначеннями та описами класів, властивостей та осіб. Правильність у цьому випадку може означати відповідність визначеним золотим стандартам. Крім того, оскільки золотий стандарт являє собою ідеальну онтологію конкретного домена, порівнюючи

вивчену онтологію з цією золотою посиланням, можна ефективно оцінити, чи добре онтологія охоплює область, і чи вона містить невідповідні елементи щодо домену.

Корпус-підходи, також відомі як підходи до управління даними, використовують, щоб оцінити, наскільки онтологія достатньо охоплює потрібну область. Концепція такого типу підходу полягає у порівнянні вивченої онтології зі змістом текстового корпусу, який суттєво охоплює цю область. Перевагою є порівняння однієї або декількох онтологій з корпусом, а не порівняння однієї онтології з іншою, яка вже існує.

Онтології для опису метрик

Робота [38] описує розроблення онтології для організації інформації про метрики та її потенційного застосування для визначення та управління метриками в проєкті CTSA (Clinical and Translational Science Award). Метою є підтримка інтегрованої бази даних всіх показників, що використовуються компонентами CTSA. Онтологія подається як концептуальна схема даних сутність–зв'язок.

Висновки і перспективи подальших досліджень

Наявні метрики можна поділити на два види:

1. Метрики, що описують сукупні інтегральні характеристики систем.

Дають змогу порівнювати різні системи між собою. Наприклад, стосовно онтологій, інтегральні характеристики – це такі характеристики онтології, які

- оцінюють онтологію загалом;
- дають можливість порівнювати різні онтології між собою.

Такими характеристиками є, наприклад, певні характеристики графу онтології:

- кількість вузлів;
- кількість ребер;
- діаметр;
- максимальний, мінімальний, середній степінь вузла тощо.

Наявні метрики цього виду не враховують суттєвих характеристик: вони (за винятком морфологічних метрик) малопридатні для порівняння різних онтологій, змодельованих за допомогою ієрархічних дерев (онтології-таксономії). Крім того, в багатьох випадках зручніше працювати зі скінченною та нормованою мірою, а за однакової ваги для всіх рівнів ієрархії дерева морфологічні метрики не забезпечують нормованості міри (а в разі нескінченного зростання кількості рівнів ієрархії – не забезпечують і скінченності міри). Причина полягає в тому, що такі метрики ігнорують вагу вузлів чи вагу зв'язків між вузлами ієрархічного дерева. Така вага є істотною у багатьох прикладних областях: деякі зв'язки важливіші, ніж інші. Зазвичай, в ієрархічних структурах на верхніх рівнях ієрархії розташовані вузли, що відповідають вагомішим сутностям. Зокрема, для онтологій – загальніші класи понять розташовані на верхніх рівнях ієрархії, для вкладених відношень – важливіші атрибути належать відношенням, які охоплюють інші, внутрішні відношення.

Отже, необхідно побудувати метрики для ієрархічних дерев, які враховуватимуть вагу рівнів ієрархії. При цьому вищі рівні ієрархії повинні мати більшу вагу, а нижчі – меншу.

2. Метрики, що описують характеристики відношень між елементами однієї системи.

Стосовно онтологій: такі метрики описують відношення між концептами, дають змогу порівнювати концепти в алгоритмах розв'язання таких завдань, як автоматичне реферування, пошук та оцінювання текстів тощо.

Всі відношення між концептами можна поділити на дві категорії:

- Таксономічні відношення.

Вони описують ієрархічні зв'язки "is-a" ("…є різновидом…") між концептами, тобто відношення "підмножина – множина", "множина – надмножина".

- Нетаксономічні відношення.

Вони описують не ієрархічні зв'язки, наприклад, зв'язок "used-in" ("термін_1 зустрічається у визначенні терміну_2") або зв'язок "uses-of" ("визначення терміну_1 використовує термін_2", чи "термін_1 посилається на термін_2").

Наявні метрики цього виду оцінюють відстань між концептами за припущення існування єдиного шляху від одного концепту до іншого.

У випадку, якщо онтологія є таксономією, тобто всі зв'язки між концептами – ієрархічні, то таке припущення справедливе і зазначені метрики правильні.

Проте, якщо існують нетаксономічні відношення між концептами (наприклад, "термін_1 посилається на термін_2"), то така метрика не буде правильною, бо вони передбачають існування багатьох шляхів від одного концепту до іншого.

Метрика, основана на багатьох зв'язках між концептами, повинна враховувати:

- відстань як кількість переходів у орієнтованому графі онтології на шляху від одного вузла-концепту до іншого;
- кількість таких шляхів.

Отже, необхідно побудувати метрику, яка дасть змогу оцінювати відстані між концептами у випадку нетаксономічних зв'язків у онтології в загальному випадку існування багатьох шляхів від одного концепту до іншого.

Список літератури

1. Колмогоров, А. Н., & Фомин, С. В. (1976). *Элементы теории функций и функционального анализа*. М., Наука.
2. Choquet, G. (1953). Theory of capacities. *Ann. Inst. Fourier (Grenoble), 5*, 31–295.
3. Denneberg, D. (1994). *Non-Additive Measure and Integral*. Dordrecht: Kluwer Academic Publishers.
4. Kimball, R. (2013). *Dimensional Modeling Techniques. Additive, Semi-Additive, and Non-Additive Facts.* Kimball group. http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/additive-semi-additive-non-additive-fact/
5. Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.
6. Blahut, Richard E. (1983). *Theory and practice of error control codes*. Addison-Wesley.
7. Grootendorst, M. (2021). *9 Distance Measures in Data Science. The advantages and pitfalls of common distance measures*. Towards data science. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa
8. Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. Vaudoise sci. Natur, 37*(140), 241–272.
9. Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab. Biol. Krifter, V*(4), 1–34.
10. Dice, Lee R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409
11. Fenton, N. E., Pfleeger, S. L. (1997). *Software Metrics: A Rigorous and Practical Approach*. International Thompson Computer Press.
12. Naylor, C. (1983). *Build your own PC expert system*. Sigma Press.
13. Теслер, Г. С. (2005). Метрики и нормы в иерархии категориальных семантик и функций. *Математичні машини і системи*, *2*, 65–68.
14. Величко, В. Ю. (2004). Розв'язання аналітичних задач в дискретних середовищах методами виведення за аналогією): дис. канд. наук, *Інститут кібернетики*, *Київ*.
15. Литвин, В. В. (2011). Бази знань інтелектуальних систем підтримки прийняття рішень. *Національний університет "Львівська політехніка"*. ISBN 978-617-607-059-7.
16. Досин, Д. Г., & Литвин, В. В. & Нікольський, Ю. В. & Пасічник, В. В. (2009). Інтелектуальні системи, базовані на онтологіях. *Цивілізація*, *Львів.*
17. Литвин, В. В. (2008). Спосіб введення метрики для визначення відстані між текстовими документами. *Інформаційні системи та мережі, 621*, 162–171.
18. Lytvyn, V. & Vysotska, V .& Dosyn, D. & Lozynska, O. & Oborska, O. (2018). Methods of building intelligent decision support systems based on adaptive ontology. *Proceedings of the IEEE Second International Conference on Data Stream Mining & Processing*.

19. Duy Hoa Ngo, & Zohra Bellahsene, & Remi Coletta (2011). A Generic Approach for Combining Linguistic and Context Profile Metrics in Ontology Matching. *ODBASE'2011: 10th International Conference on Ontologies, DataBases, and Applications of Semantics, Oct 2011, Crete, Greece*, 800–807.

20. Alsayed Algergawy, & Samira Babalou, & Birgitta Konig-Ries (2016). A New Metric To Evaluate Ontology Modularization. *2nd International Workshop on Summarizing and Presenting Entities and Ontologies Co-located with the 13th Extended Semantic Web Conference*. *Greece, 2016-05-30*. http://ceur-ws.org/Vol-1605/paper4.pdf.

21. Giorgos Stoilos, & Giorgos Stamou, & Stefanos Kollias (2005). A String Metric for Ontology Alignment. *International Semantic Web Conference ISWC 2005: The Semantic Web – ISWC*, 624–637.

22. García, J. & García-Peñalvo, F. J. & Therón, R. (2010). A Survey on Ontology Metrics. *World Summit on Knowledge Society WSKS 2010: Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, 22–27.

23. Denny Vrandecic, & York Sure (2007). How to Design Better Ontology Metrics. In *The Semantic Web: Research and Applications*, 311–325, Springer-Berlag.

24. Harith Alani, & Christopher Brewster, & Nigel Shadbolt (2006). Ranking Ontologies with AKTiveRank. *Proceedings of the International Semantic Web Conference, ISWC, 2006 5th International Semantic Web Conference (ISWC), November 2006*, Georgia, USA

25. Harith Alani, & Christopher Brewster (2006). Metrics for Ranking Ontologies. *4th Int. EON Workshop, 15th Int. World Wide Web Conference*.

26. Anthony Orme, & Haining Yao, & Letha Etzkorn. (2006). Coupling Metrics for Ontology-Based Systems. *IEEE Software*, 102–108.

27. Haining Yao, & Anthony Orme, & Letha Etzkorn. (2005). Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science*, *1*(1), 107–113.

28. Yinglong Ma, & Beihong Jin, & Yulin Feng (2009). Semantic oriented ontology cohesion metrics for ontology-based systems. *The Journal of Systems and Software*, *Elsevier*

29. Nicola Guarino, & Chris Welty (2004). An Overview of OntoClean. *The Handbook on Ontologies*. *Berlin: Springer-Verlag*, 151–172.

30. Nicola Guarino, & Chris Welty (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM, ACM Press*, 61–65,.

31. Zhe YANG, & Dalu Zhang, & Chuan Ye. (2006). Evaluation Metrics for Ontology Complexity and Evolution Analysis. *IEEE International Conference on e-Business Engineering (ICEBE'06)*.

32. Joe Raad, & Christophe Cruz (2015). A Survey on Ontology Evaluation Methods. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, At Lisbon, Portugal, November 2015*.

33. Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web* (pp. 251–263). Springer Berlin Heidelberg.

34. Ponzetto, S. P., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. *AAAI*, *7*, 1440–1445.

35. Treeratpituk, P., & Khabsa, M., & Giles, C. L. (2013). Graph-based Approach to Automatic Taxonomy Generation (GraBTax). *arXiv preprint arXiv:1307.1718*.

36. Elias Zavitsanos, & George Paliouras, & George A. Vouros (2011). Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Trans. on Knowl. and Data Eng.*, *23*(11), 1635–1648.

37. Kashyap, V., & Ramakrishnan, & C., Thomas, C., & Sheth, A. (2005). TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, *1*(2), 240–266.

38. Dagobert Soergel, & Olivia Helfer (2016). A Metrics Ontology. An intellectual infrastructure for defining, managing, and applying metrics. *Knowl Organ Sustain World Chall Perspect Cult Sci Technol Shar Connect Soc*., *15*, 333–341.

References

1. Kolmogorov A. N., & Fomin S. V. (1976). *Elements of the theory of functions and functional analysis*. Nauka (Science).

2. Choquet, G. (1953). Theory of capacities. *Ann. Inst. Fourier (Grenoble)*, *5*, 31–295.

3. Denneberg, D. (1994). *Non-Additive Measure and Integral*. Dordrecht: Kluwer Academic Publishers.

4. Kimball, R. (2013). *Dimensional Modeling Techniques. Additive, Semi-Additive, and Non-Additive Facts.* Kimball group. http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/additive-semi-additive-non-additive-fact/

5. Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.

6. Blahut, R. E. (1983). *Theory and practice of error control codes*. Addison-Wesley.

7. Grootendorst, M. (2021). *9 Distance Measures in Data Science. The advantages and pitfalls of common distance measures*. Towards data science. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

8. Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. Vaudoise sci. Natur*, *37*(140), 241–272.

9. Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab. Biol. Krifter*, *V*(4), 1–34.

10. Dice, Lee R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409

11. Fenton, N. E., Pfleeger, S. L. (1997). *Software Metrics: A Rigorous and Practical Approach*. International Thompson Computer Press.

12. Naylor, C. (1983). *Build your own PC expert system*. Sigma Press.

13. Tesler, G. S. (2005). Metrics and norms in the hierarchy of categorical semantics and functions. *Mathematical machines and systems*, *2*, 65–68.

14. Velychko, V. Yu. (2004). Solving analytical problems in discrete media by inference methods by analogy.

15. Lytvyn, V. (2011). Knowledge base of intelligent decision support systems (original title: В. В. Литвин. Бази знань інтелектуальних систем підтримки прийняття рішень). *Lviv Polytechnic Publishing House*. ISBN 978-617-607-059-7.

16. Dosyn, D. & Lytvyn, V. & Nikolsky,Yu. & Pasichnyk, V. (2009). Intelligent systems based on ontologies.

17. Lytvyn, V. (2008). A method of entering metrics to determine the distance between text documents. *Information systems and networks*, *621*, 162–171.

18. Lytvyn,V. & Vysotska, V. & Dosyn, D. & Lozynska,O. & Oborska, O. (2018). Methods of building intelligent decision support systems based on adaptive ontology. *Proceedings of the IEEE Second International Conference on Data Stream Mining & Processing*.

19. Duy Hoa Ngo, & Zohra Bellahsene, & Remi Coletta. (2011). A Generic Approach for Combining Linguistic and Context Profile Metrics in Ontology Matching. *ODBASE'2011: 10th International Conference on Ontologies, DataBases, and Applications of Semantics, Oct 2011, Crete, Greece*, 800–807.

20. Alsayed Algergawy, & Samira Babalou, & Birgitta Konig-Ries. (2016). A New Metric To Evaluate Ontology Modularization. *2nd International Workshop on Summarizing and Presenting Entities and Ontologies Co-located with the 13th Extended Semantic Web Conference. Greece, 2016-05-30*. http://ceur-ws.org/Vol-1605/paper4.pdf.

21. Giorgos Stoilos, & Giorgos Stamou, & Stefanos Kollias. (2005). A String Metric for Ontology Alignment. *International Semantic Web Conference ISWC 2005: The Semantic Web – ISWC*, 624–637.

22. García, J. & García-Peñalvo, F. J., & Therón, R. (2010). A Survey on Ontology Metrics. *World Summit on Knowledge Society WSKS 2010: Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, 22–27.

23. Denny Vrandecic, & York Sure. (2007). How to Design Better Ontology Metrics. In *The Semantic Web: Research and Applications*, 311–325, Springer-Berlag.

24. Harith Alani, & Christopher Brewster, & Nigel Shadbolt (2006). Ranking Ontologies with AKTiveRank. *Proceedings of the International Semantic Web Conference, ISWC, 2006 5th International Semantic Web Conference (ISWC), November 2006*, Georgia, USA

25. Harith Alani, & Christopher Brewster (2006). Metrics for Ranking Ontologies. *4th Int. EON Workshop, 15th Int. World Wide Web Conference*.

26. Anthony Orme, & Haining Yao, & Letha Etzkorn (2006). Coupling Metrics for Ontology-Based Systems. *IEEE Software*, 102–108.

27. Haining Yao, & Anthony Orme, & Letha Etzkorn (2005). Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science*, *1*(1), 107–113.

28. Yinglong Ma, & Beihong Jin, & Yulin Feng (2009). Semantic oriented ontology cohesion metrics for ontology-based systems. *The Journal of Systems and Software*, *Elsevier*

29. Nicola Guarino, & Chris Welty (2004). An Overview of OntoClean. *The Handbook on Ontologies*. *Berlin: Springer-Verlag*, 151–172.

30. Nicola Guarino, & Chris Welty (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM, ACM Press*, 61–65.

31. Zhe YANG, & Dalu Zhang, & Chuan YE. (2006). Evaluation Metrics for Ontology Complexity and Evolution Analysis. *IEEE International Conference on e-Business Engineering (ICEBE'06)*.

32. Joe Raad, & Christophe Cruz (2015). A Survey on Ontology Evaluation Methods. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, At Lisbon, Portugal, November 2015*.

33. Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web*, 251–263. Springer Berlin Heidelberg.

34. Ponzetto, S. P., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. *AAAI*, *7*, 1440–1445.

35. Treeratpituk, P., & Khabsa, M., & Giles, C. L. (2013). Graph-based Approach to Automatic Taxonomy Generation (GraBTax). *arXiv preprint arXiv:1307.1718*.

36. Elias Zavitsanos, & George Paliouras, & George A. Vouros. (2011). Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Trans. on Knowl. and Data Eng.*, *23*(11),1635–1648.

37. Kashyap, V., & Ramakrishnan, & C., Thomas, C., & Sheth, A. (2005). TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, *1*(2), 240–266.

38. Dagobert Soergel, & Olivia Helfer (2016). A Metrics Ontology. An intellectual infrastructure for defining, managing, and applying metrics. *Knowl Organ Sustain World Chall Perspect Cult Sci Technol Shar Connect Soc*., *15*, 333–341.

ANALYSIS OF METRICS FOR INTELLIGENT INFORMATION SYSTEMS

Viktor Hryhorovych,
Lviv Polytechnic National University
viktor.grigorovich@gmail.com, 0000-0002-5828-067X

The problem of constructing metrics is crucial for solving the problem of quantitative evaluation of both systems of objects of arbitrary nature as a whole and the relationships that describe the connections between the components of these systems. Modern information systems simulate subject areas that contain objects and systems of complex structure. The network model is most appropriate for describing the world around it: it reflects objects and systems of objects of arbitrary nature that interact with each other. In fact, any system can be described using a network model. Hierarchical models should be singled out as a kind of network models of complex systems. Hierarchical models are very widespread and are used in various fields – in biology, sociology, economics, technology, management, etc. - each industry has a set of its own hierarchical models. The paper analyzes metrics suitable for evaluating intelligent information systems, in particular – systems that are based on ontologies, non-relational (hierarchical) databases, non-normalized (nested) relationships.

Key words: metrics, intelligent information system, ontology, network model, hierarchical model.