УДК 004.94 https://doi.org/10.15407/pp2021.03.016

О.В. Захарова

ВИЗНАЧЕННЯ СТУПЕНЯ СЕМАНТИЧНОЇ ПОДІБНОСТІ З ВИКОРИСТАННЯМ АПАРАТУ ДЕСКРИПТИВНИХ ЛОГІК

Встановлення семантичної подібності інформації є невід'ємою складовою процесу вирішення будь – яких задач інформаційного пошуку, в тому числі задач, пов'язаних з обробкою великих даних, виявленням семантичних веб сервісів, категоризації та класифікації інформації тощо. Введення спеціальних функцій для визначення кількісних показників ступеня семантичної відповідності інформації дозволяють ранжувати знайдену інформацію за її семантичною близькістю до цілі або пошукового запиту/ шаблону. Формування таких оцінок повинно враховувати багато аспектів: від сутності самих понять, семантична близькість яких підлягає оцінюванню, до особливостей бізнес-задачі, в межах вирішення якої це робиться. Зазвичай при побудові таких функцій подібності семантичні підходи поєднуються зі структурними, що забезпечують синтаксичне порівняння описів концептів. Це дозволяє суттєво деталізувати сам опис концепту, а вплив синтаксичної відповідності можна значно зменшити, використовуючи для представлення інформації виразніші дескриптивні логіки (ДЛ) та шляхом перенесення фокусу на семантичні властивості. ДЛ-онтології, на сьогодні, є найбільш розвинутим засобом представлення семантики, а механізми міркувань ДЛ забезпечують можливість логічного висновку. Більшість наведених у статті оцінок будуються на основі базових ДЛ, що підтримують лише констуктор перетину, але описані підходи можуть бути застосовані для будь-якої ДЛ, що забезпечує базові сервіси міркувань.

У роботі проведений аналіз існуючих підходів, моделей та способів оцінювання, заснованих на застосуванні апарату дескриптивних логік, запропонована їхня класифікація як за рівнем визначення подібності, так і за видами співставлення. Основна увага приділяється встановленню подібності концептів (моделям понятійного рівня). Задачі встановлення ступеня подібності між екземплярами та між концептом і екземпляром зводяться до знаходження найбільш специфічного концепту для екземпляра/екземплярів та оцінювання подібності відповідних концептів. Уведено поняття екзистенційної подібності та продемонстровано застосування певних видів оцінок для визначення ступеня семантичної подібності понять та/або знань на прикладі онтології геометричних понять.

Ключові слова: семантична подібність інформації, ступінь подібності концептів, найменше спільне покриття, оцінки вимірювання подібності, найбільш специфічний концепт, найбільш специфічний попередник, функція подібності, подібність за інформаційним змістом, семантична подібність за відповідністю ознак, функція відстані шляху, моделі оцінювання на основі властивостей, моделі оцінювання на основі семантичної мережі, моделі оцінювання на основі інформаційного контенту, екзистенційна подібність концептів, подібність екземплярів, подібність концепту та екземпляра, подібність ДЛ описів, GCS-подібність.

Вступ

Проблема знаходження семантично схожих понять та визначення ступеня їхньої подібності є нагальною як для вирішення прикладних задач (виявлення семантичних сервісів, ефективного семантичного пошуку інформації, категоризації даних тощо), так і більш загальних задач інформаційних технологій, як, наприклад, інтеграція онтологій, знань, інформаційний пошук тощо. Існує багато підходів, які намагаються вирішити проблеми визначення подібності методами текстового аналізу або за допомогою використання спеціальних словників понять, зокрема, словника Wordnet [1]. Водночас, зазвичай, розглядаються лише атомарні концепти, а складніші лишаються поза увагою. Окрім цього, опускаються варіанти виявлення подібності серед екземплярів понять та подібності екземпляра та концепту. Також, слід зазначити, що ступені інформаційної подібності повинні базуватися на семантиці, оскільки суто синтаксичний підхід є надто слабким, щоб забезпечити виконання стандартних виведень, особливо, якщо взяти до уваги виразні дескриптивні логіки (зокрема, $\mathcal{ALC}$), як засіб представлення знань. Зрозуміло, що алгоритми визначення та функції мір відповідності повинні бути ефективними. Якщо вони будуть надто складними, то навряд чи зможуть забезпечити потрібний результат у допустимий період часу та стати загально-використовуваними.

Останнім часом з'явилося чимало досліджень, у яких наголошується на доцільності

використання онтологій та функцій семантичної подібності на їхній основі для порівняння концептів та/або екземплярів концептів, котрі можна отримати через інтеграцію гетерогенних джерел інформації [2,3,4,5].

Види та рівні визначення подібності

Подібність інформації/знань можна розглядати та визначати на різних рівнях. А саме можна виділити:

1) понятійний рівнь – визначення подібності концептів;

2) рівень знань – визначення подібності екземплярів концептів;

3) змішаний рівень – визначення подібності концепту та екземпляра.

Оцінки, що вимірюють подібність, зазвичай використовують базову теорію множин та базуються на спільності об'єктів. Зокрема, базовий критерій для визначення таких мір можна сформулювати наступним чином: значення подібності між об'єктами є не лише результатом їх спільних характеристик, а й результатом їхніх відмінностей. Цей критерий відповідає теоретико-інформаційному визначенню подібності, а об'єктами в даному випадку є концепти та екземпляри концептів.

Розглянемо підходи до визначення можливих оцінок ступеня подібності та відповідні моделі оцінювання для кожного з перелічених рівнів. Але спочатку визначимо ряд понять, що використовуються більшістю існуючих моделей оцінювання.

Базові поняття та визначення

Визначення 1. **LCS (Least Concept Subsumer)** [23, 24] - найменше спільне покриття концептів. Нехай L дескриптивна логіка (ДЛ). Опис концепту E дескриптивної логіки $\mathcal{L}$ є найменшим загальним покриттям (LCS) описів концептів $C_1, \cdots, C_n$ в $\mathcal{L}$ (скорочено LCS($C_1, \cdots, C_n$)), якщо:

1) $C_i \sqsubseteq E$ для $i = 1, \cdots, n$ та

2) E є найменшим описом $\mathcal{L}$-концептом, що задовільняє (1), тобто, якщо E_0 є описом L-концепту такий, що $C_i \sqsubseteq E_0$ для всіх $i = 1, \cdots, n$, то $E \sqsubseteq E_0$.

Одразу слід зазначити, що LCS існує не для будь-якої ДЛ, що використовується для представлення знань. Але, якщо LCS існує, то воно є унікальним з точністю до еквівалентності. Всі міри, що розглядатимуться нижче, базуються на базовій ДЛ $\mathcal{ALC}$. У [6] показано, що для $\mathcal{ALC}$ логіки LCS існує завжди та задається диз'юнкцією понять. У разі, якщо диз'юнкція не підтримується логікою, LCS обчислюється вибором загальних імен понять в описах концептів (у межах понять універсуму та екзистенційних обмежень для тієї самої ролі), не зважаючи на TBox в цілому [6]. Але в такому разі результат обчислення LCS може бути надто загальним. За цими міркуваннями LCS обчислюється відносно TBox, на основі якого визначаються концепти [7].

Беручи до уваги TBox, визначення LCS можна переформулювати так.

Визначення 2. Нехай $\mathcal{L}1$ та $\mathcal{L}2$ дескриптивні логіки такі, що $\mathcal{L}1$ є під-логікою $\mathcal{L}2$, тобто $\mathcal{L}1$ містить менше конструкторів, які використовуються для побудови виразів. Для заданого TBox логіки $\mathcal{L}2$ $\mathcal{T}$, $\mathcal{L}1(\mathcal{T})$ – множина описів концептів, які можуть містити концепти, що визначені у $\mathcal{T}$. $C_1, \ldots, C_n$ описи концептів з $\mathcal{L}1(\mathcal{T})$, тоді LCS($C_1, \ldots, C_n$) у $\mathcal{L}1(\mathcal{T})$ відносно TBox $\mathcal{T}$ є опис найбільш специфічного $\mathcal{L}1(\mathcal{T})$ концепту, який включає $C_1, \ldots, C_n$ на TBox $\mathcal{T}$, а саме, це такий опис $\mathcal{L}1(\mathcal{T})$- концепту D, що:

3) C_i D для $i = 1, \cdots, n$ та

4) Якщо E є описом $\mathcal{L}1(\mathcal{T})$- концепту, що задовільняє C_i E для всіх $i = 1, \cdots, n$, то $D \sqsubseteq E$

Якщо LCS відповідно для TBox не існує (наприклад, у разі циклічного TBox), обчислюється його апроксимація, що називається Гарне Загальне Покриття (GCS) [25] відносно TBox та існує навіть для загального TBox. GCS обчислюється через визначення найменшої кон'юнкції концептів та їхніх заперечень, що може включати кон'юнкцію імен концептів вищого рівня для кожного з концептів, що розглядається, та аналогічної кон'юнкції концептів, які становлять ранг екзистенційних та універсальних обмежень відносно тієї самої ролі. GCS є специфічнішим покриттям, ніж LCS, що обчислюється безвідносно Tbox. Хоча у загальному випадку воно включає (або є еквівалентним) LCS, що обчислюється відносно TBox [7].

MSA (Most Specific is-a Ancestor) [8] – найбільш специфічний попередник в ієрархії таксономії. Дане поняття визначається як бінарне відношення на таксономії концептів, але за семантикою воно подібне до LCS. Ці обидва поняття обчислюють найспецифічніше уза-

гальненя вхідних концептів (відносно операції включення). Різниця ж полягає в тому, що MSA працює на таксономії понять та повертає один концепт, який містить два вихідних концепти (є їхнім is-a попередником) та не включає жодного іншого, який би задовільняв ті ж вимоги. A LCS є описом, який покриває вихідні концепти, та, як результат, при його обчисленні повертаються всі включені до нього концепти. Якщо концепти зв'язані лише родо-видовими зв'язками, тобто TBox є таксономією, LCS покриття концептів вироджується до одного попередника і $LCS(C_1,C_2)=MSA(C_1,C_2)$.

MSC - найбільш специфічний концепт. Унарне відношення на множині екземплярів ABox.

Визначення 3. [25] Нехай дано ABox $\mathcal{A}$ та екземпляр α цього ABox, найбільш специфічним концептом для екземпляра a відносно ABox $\mathcal{A}$ є концепт C, позначається , такий, що $\mathcal{A}C()$, та такого, що $\mathcal{A}\models D()$, $C\sqsubseteq D$ (де $\models$ позначає оператор виведення).

Одразу слід зазначити, що в загальному випадку ациклічного ABox у виразній ДЛ не може бути виражений кінцевим описом концепту [2], можна отримати лише його апроксимацію. Тож, існування найбільш специфічного концепту для індивіда ABox не є гарантованим, або його складно обчислити, й апроксимацію обмежують певною встановленою глибиною. Максимальна глибина апроксимації, як визначено у [20], відповідає глибині ABox. У такому разі, для будь - якого екземпляра ABox α можемо визначити найбільш специфічний концепт MSC(α) або його апроксимацію MSC*(α).

Визначення семантичної подібності концептів

На сьогодні існує чимало досліджень, автори яких намагаються перевести семантичні відношення між поняттями у деякі кількісні показники. Зрозуміло, що на принципи формування таких оцінок впливає, насамперед, сутність самих понять, семантична близькість яких оцінюється, а також задача, для вирішення якої обираються чи визначаються функції подібності. Більшість існуючих досліджень застосовують семантичний підхід у поєднанні зі структурним, який порівнює описи концептів, що розглядаються. Звісно, це дозволяє суттєво деталізувати опис, а вплив синтаксичної відповідності можна значно зменшити при використанні для представлення інформації більш виразних ДЛ та перенесення фокусу на семантичні властивості.

При встановленні ступеня семантичної відповідності між концептами однієї онтології функція подібності фактично є відображенням $\mathcal{S}$: $\mathcal{L}(\mathcal{T})\times\mathcal{L}(\mathcal{T})\rightarrow Y$, де $\mathcal{T}$ є TBox даної онтології, представлений у ДЛ $\mathcal{L}$, а Y є дійсним числом, що кількісно визначає ступінь подібності. В оцінках, що базуються на співвідношеннях (частках), $Y\in[0,1]$, але існують й інші моделі вимірювань.

У загальному випадку задача суттєво ускладнюється. Якщо відповідність встановлюється між концептами двох різних онтологій з TBox-амі $\mathcal{T}1$ та $\mathcal{T}2$ ДЛ, відповідно $\mathcal{L}1$ та $\mathcal{L}2$, необхідно побудувати відображення $\mathcal{S}$: $\mathcal{L}1(\mathcal{T}1)\times\mathcal{L}2(\mathcal{T}2)\rightarrow Y$.

У будь-якому випадку функція повинна мати наступні властивості:

1) Нехай E – певна множина елементів (об'єктів однієї чи різних онтологій), для яких визначається ступінь подібності, то функція $\mathcal{S}$ визначена на множині $E\times E$

2) Функція $\mathcal{S}$ є позитивно-визначеною, тобто $\mathcal{S}(C,D)\geq 0$

3) $\forall C,D: \mathcal{S}(C,D)\leq\mathcal{S}(C,C)$

При визначенні функції відповідності, необхідно розуміти, що подібність концептів можна розглядати як з боку ступеня їх подібності, так і ступеня їх відмінності. І функція подібності повинна мати позитивну кореляцію зі ступенем подібності між концептами та негативну - з показником відмінності між ними. Зрозуміло, що цей числовий показник залежить від багатьох факторів, а саме: специфіки досліджуваного контенту, виразності та однорідності мов представлення онтологій тощо. Але ключовим питанням при створенні функції подібності є «як виміряти ступінь подібності (відмінності) концептів». Це, в свою чергу, пов'язано з тим, як збирається досліджувана інформація. Навряд чи показник подібності можна розцінювати, як абсолютну оцінку, але він має забезпечувати можливість достовірного ранжування концептів за ступенем їхньої подібності. Серед основних підходів до побудови такої функції можна виділити:

1) визначення подібності як функції відстані шляху між таксонами в ієрархії, що лежить в основі цієї онтології [10, 11, 12];

2) оцінка семантичної подібності за відповідністю ознак [13];

3) визначення ступеня подібності концептів за інформаційним змістом [14,15];

4) екзістенціональна подібність понять.

Перший підхід може бути застосований лише в межах однієї онтології, тобто його використання може бути доцільним лише, якщо оцінювання виконується на базі одного джерела інформації, і досліджувані поняття, є концептами однієї онтології або інтегрованої онтології вихідних джерел інформації. Другий підхід для обчислення семантичної подібності використовує як загальні, так і дискримінантні ознаки між поняттями та / або екземплярами понять. Методи третьої групи засновані на теорії інформації. Вони визначають ступінь подібності між двома поняттями в рамках ієрархії понять з точки зору кількості інформації, що передається безпосередньо супер-концептом, який включає порівнювані концепти. Всі оцінки, які базуються на ознаках та властивостях концептів, можна назвати оцінками інтенсіональної подібності. Під *екзистенційною подібністю понять* будемо розуміти ступінь їхньої близькості за множинами екземплярів, які вони містять.

У разі встановлення ступеня відповідності між поняттями різних, можливо, різнорідних, онтологій, перелічені вище підходи працюють за умови виконання певних чинників та обмежень. По-перше, формальне представлення цих онтологій повинна підтримувати механізми міркувань, такі як включення. (Слід одразу зазначити, що механізм включення підтримується базовими ДЛ такими, зокрема, як $\mathcal{ALC}$). По-друге, застосування підходів оцінювання базується на використанні узагальненої онтології, а локальні концепти в різних онтологіях повинні успадковувати структуру визначення з їх узагальненої онтології. У [16] пропонуються деякі підходи до порівняння таких концептів із різних онтологій за складом їхніх екземплярів. А саме, робиться припущення, що при виконанні визначених вище обмежень, ознакою відповідності двох концептів може бути перетин множин їхніх екземплярів. А для порівняння описів понять, які можна поєднати у загальну онтологію, використовуються три основні підходи, а саме:

- фільтрація на основі відстані-шляху між поняттями у загальній онтології;
- визначення ступеня подібності на основі відповідності елементів графів (один до одного) описів понять;
- визначення ймовірносних метрик, які визначають подібність з точки зору спільного розподілу понять.

Також, якщо обчислення оцінок подібності робиться для концептів з різних онтологій, необхідно враховувати різницю між рівнями формалізації специфікацій цих онтологій. Зокрема, у [17] функція відповідності визначає класи подібних сутностей за допомогою співставлення з використанням наборів синонімів, семантичного сусідства та дискримінаційних ознак, що класифіковані за частинами, функціями та атрибутами. У [9] представлений інший підхід, спрямований на знаходження спільних властивостей серед концептів або тверджень.

Перелічені групи підходів до оцінювання подібності базуються на відповідних моделях.

Основні моделі оцінювання

До найбільш розповсюджених моделей оцінювання можна віднести:

- моделі на основі властивостей;
- моделі на основі семантичних мереж;
- моделі на основі інформаційного контенту.

В **моделях на основі властивостей концепт** C характеризується множиною своїх властивостей, що позначається ftrs(C). У [18] пропонується дві групи вимірювань для такої моделі:

1) контрастна модель, де подібність двох концептів C і D визначається лінійною функцією

$$contra(C,D) = \theta f\big(ftrs(C) \cap ftrs(D)\big) - \alpha f(ftrs(C) \backslash ftrs(D)) - \beta f(ftrs(D) \backslash ftrs(C)),$$

де \ - операція різниці множин, α, β та θ не негативні константи, а f(.) – виражає кількість ознак в множині

2) нормалізована модель співвідношення, де подібність визначається як частка множин:

$$sim(C,D) =_{def} \frac{f(ftrs(C) \cap ftrs(D))}{f\big(ftrs(C) \cap ftrs(D)\big) + \alpha f\big(ftrs(C) \backslash ftrs(D)\big) + \beta f(ftrs(D) \backslash ftrs(C))}$$

Якщо вважати, що функція подібності є симетричною, то Якщо припустити, що функція є дистрибутивною по множинам, що не перетинаються, можна перетворити наступним чином:

$$sim(C,D) =_{def} \frac{2f(ftrs(C) \cap ftrs(D))}{f(ftrs(C)) + f(ftrs(D))}$$

У **моделях, що засновані на семантичній мережі**, довідкова інформація надається у формі семантичної мережі, що включає концепти та, принаймні, is-a ребра (іноді розглядаються більш складні відносини, як у WordNet). Це є прикладом саме того випадку, коли оцінювання подібності базується на вимірюванні довжини шляху між концептами у мережі. Якщо концепти знаходяться у таксономії, тобто пов'язані родовидовими відношеннями, значення подібності між двома концептами обчислюються кількістю ребер на шляху від концептів, що розглядаються, до їх найближчого попередника. Якщо поняття розділені лише декількома зв'язками, то вони вважаються подібними. Чим більше зв'язків їх розділяють, тим менша схожість між ними [8, 19, 12, 20]. Тобто, для оцінювання відповідності концептів C і D знаходиться найбільш специфічний is-a попередник E = MSA(C,D) концептів C і D та обчислюється міра подібності як сума довжин шляхів від C до E та від E до D. Розвинутіші оцінки можуть враховувати глибину концепту MSA(C, D), щільність ребер у вузлах шляху та вагу ребер.

У **моделях, заснованих на інформаційному контенті**, разом із семантичною мережею використовується інформація про імовірність того, що сутність описується конкретним концептом C. Така імовірність зазвичай оцінюється на основі вихідної специфічної задачі.

Величина інформаційного контенту концепту вимірюється на основі імовірності $pr(\mathrm{C})$, як $IC(\mathrm{C}) =_{def} -\log pr(\mathrm{C})$. У [21] пропонується міра подібності концептів C та D на основі імовірносної оцінки їх MSA:

$$sim(C,D) =_{def} IC(MSA(C,D)) =_{def} -\log pr(MSA(C,D)).$$

У [22] пропонується міра відстані концептів у мережі на основі їхнього інформаційного контенту, що враховує такі фактори, як глибина та щільність ребер шляху між концептами:

$$dist(C,D) =_{def} IC(C) + IC(D) - 2IC(MSA(C,D))$$

У [18] пропонується міра подібності, що визначається часткою:

$$sim(C,D) =_{def} \frac{2IC(MSA(C,D))}{IC(C) + IC(D)}$$

Визначення оцінювання подібності для ДЛ описів

Усі метрики, представлені вище, визначені на атомарних концептах. Але наведені оцінки можна переформулювати й для складних концептів, які визначаються через атомарні засобами ДЛ. Зазначимо, що при побудові оцінок ми вважаємо, що описи концептів представлені у базовій ДЛ, яка підтримує лише операцію перетину концептів. Будь-який опис концепту можна привести до його нормальної форми, тобто розкласти так, щоб він містив лише атомарні концепти. Зазвичай, це робиться просто шляхом підстановки у визначеня замість неатомарних концептів їхні описи. Позначимо через nf(C) множину атомарних концептів, що зустрічаються у нормальній формі концепта C. Зазначимо, що C⊑D (де ⊑ - просте структурне включення), якщо nf(D) ⊆ nf(C).

Із урахуванням наведеного визначення струкурного опису концепту можна переформулювати наведені вище оцінки відповідності наступним чином.

Для моделі властивостей будемо розглядати властивості концепту як атомарні концепти, а складний концепт як кон'юнкцію цих атомарних концептів. Враховуючи особливості перетину та різниці множин атомарних властивостей, міри відповідності, за умови їхньої симетричності, можна визначити так:

$$contra(C,D) =_{def} f(lcs(C,D)) - 0{,}5 * * f(diff(C,D)) - 0{,}5 * f(diff(D,C))$$

$$sim(C,D) =_{def} \frac{2 * f((C,D))}{2 * f(lsc(C,D)) + f(diff(C,D) + f(diff(D,C))}$$

Слід зазначити, що у наведених мірах функція f є лічильником властивостей, можливо зважених.

Наразі розглянемо модель семантичної мережі. Якщо мережа є ієрархією, та

концепт С має is-a попередників: U_1, U_2, U_3, …, U_{n-1}, Un, введемо концепт С* такий, що $C := C^*U_1U_2U_3...U_{n-1}U_n$. У результуючому Т-Box, концепт, що визначається, має таку саме ієрархію включення як вихідні вузли у семантичній мережі, більше того, якщо у вихідній мережі шлях U_1, U_2, . . . U_n = до кореня is-a ієрархії, то нормальна форма концепта U_1 у ДЛ $nf(U_1)$ = . Іншими словами, якщо мережа є деревом, то кардинальність концепту, що є нормальною формою концепту С $|nf(C)|$ дорівнює довжині шляху від U_1 до кореня. Шляхи від концептів С та D до кореня перетинаються в $E=MSA(C,D)$, що на ієрархії включення співпадає з $LCS(C,D)$. Тоді відстань між концептами С і D можна визначити, як:

$$dist(C,D) =_{def} |nf(C)| + |nf(D)| - 2 * |nf(LCS(C,D))|,$$ де |X| - кардинальність концепту X.

Відповідно для інформаційних моделей:

$$dist(C,D) =_{def} IC(C) + IC(D) - 2 * IC(lcs(C,D)),$$

$$sim(C,D) =_{def} \frac{2 * IC(lsc(C,D))}{IC(C) + IC(D)}$$

Екзистенційні оцінки подібності концептів

В екзистенційних підходах значення подібності обчислюється шляхом підрахунку спільних екземплярів розширень концептів [26] або шляхом вимірювання варіації змісту між концептами, що розглядається у [27, 28, 29].

Зазвичай, онтологія має структуру, складнішу за просту таксономію, і оцінки подібності, що базуються на відстанях в таксономії або на використанні поняття найбільш специфічного попередника (MSA), використовуватися не можуть.

Слід зазначити, що семантичне відношення включення базується на канонічній інтерпретації ABox дескриптивної логіки та припущенні унікального простору імен (UNA), з якої випливає, що інтерпретацією екземплярів ABox є вони самі, а також індивіди, що відповідають різним об'єктам предметної області, мають різні імена в просторі імен. Тому визначимо оцінку відповідності концептів на основі їх розширення у каноничній інтепретації ДЛ [25].

Нехай – множина концептів у ДЛ $\mathcal{ALC}$, а $\mathcal{A}$ – ABox з каноничною інтерпретацією $\mathcal{I}$. Семантична подібність концептів $\mathcal{s}$ є функцією: $\mathcal{s}$:$\mathcal{L}\times\mathcal{L}\rightarrow$[0,1], що визначається наступним чином:

$$\mathcal{s}(C,D) = \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|+|D^{\mathcal{I}}|-|I^{\mathcal{I}}|} * \max(|I^{\mathcal{I}}|/|C^{\mathcal{I}}|, |I^{\mathcal{I}}|/|D^{\mathcal{I}}|),$$ де $I = C \sqcap D$ та $(.)^{\mathcal{I}}$ розширення концепта в інтерпретації $\mathcal{I}$.

Наведену оцінку можна обгрунтувати так. Якщо концепти С та D еквівалентні, тобто C⊑D та D⊑C, $\mathcal{s}$=1. Якщо концепти є взагалі різними, та їхні розширення не перетинаються, оцінка є мінімальною, тобто її значення дорівнює 0. У випадку *непустого перетину концептів оцінка набуває* значення в діапазоні від 0 до 1. Тобто дана оцінка виражає ступінь подібності концептів С та D, зменшену на $\max(|I^{\mathcal{I}}|/|C^{\mathcal{I}}|, |I^{\mathcal{I}}|/|D^{\mathcal{I}}|)$, що, у свою чергу, представляє несхожість цих концептів. Це означає, що ступінь подібності розглядається не як абсолютна величина, а як зважена щодо ступеня несхожесті. Така оцінка відповідає досить міцному семантичному зв'язку між поняттями, що забезпечується відношенням включення.

Оцінки GCS-подібності концептів

Міри GCS-подібності визначаються на основі поняття GCS-покриття та можуть бути застосовані, якщо оцінки, які базуються на перекритті розширень концептів, інформаційному контенті чи на відстанях між концептами, не працюють. Оцінки на основі GCS також використовують поняття розширення концепту, але замість підрахунку спільних екземплярів даних концептів, значення подібності визначається як варіація числа екземплярів у розширенні концепту відносно числа екземплярів у розширенні їх загального супер-концепту. Загальний супер-концепт визначається через GCS концептів, а оцінка відносно TBox $\mathcal{T}$ логіки $\mathcal{ALC}$ формально визначається таким чином.

$\mathcal{T}$ - $\mathcal{ALC}$- TBox. $\mathcal{L}$ – дескриптивна логіка, що включає $\mathcal{ALC}$. С і D описи концептів в $\mathcal{L}(\mathcal{T})$. Тоді міра семантичної подібності $\mathcal{S}$ є функцією $\mathcal{S}$: $\mathcal{L}(\mathcal{T}) \times \mathcal{L}(\mathcal{T}) \rightarrow$[0,1], що визначається так:

$$\mathcal{S}(\mathrm{C},D)=\frac{\min\left(|C^{\mathcal{J}}|,|D^{\mathcal{J}}|\right)}{|(GCS(C,D))^{\mathcal{J}}|}*$$

$$*\left(1-\frac{\left|(GCS(C,D))^{\mathcal{J}}\right|}{|\Delta^{\mathcal{J}}|}*\left(1-\frac{\min(|C^{\mathcal{J}}|,|D^{\mathcal{J}}|)}{\left|(GCS(C,D))^{\mathcal{J}}\right|}\right)\right),$$

де – обчислює розширення концепту відносно інтерпретації $\mathcal{J}$ (канонічної інтерпретації [2, 9]).

Тобто, якщо два концепти семантично подібні, вони повинні мати гарний спільний супер-концепт, що є близьким до обох концептів, а саме розширення суперконцепту, що містить багато екземплярів, спільних з вихідними концептами. В такому разі значення функції буде наближатися до одиниці. Навпаки, якщо вихідні концепти дуже різні, то їхні GCS та суперконцепт містить багато екземплярів, що не належать вихідним концептам, тобто значення оцінки подібності буде наближатися до 0. Дана міра не вимагає перетину концептів, що розглядаються, та не бере до уваги відстань семантичного шляху. Більше того, щоб запобігти отриманню некоректного значення подібності тоді, коли один концепт є дуже подібним до супер-концепту та дуже відрізняється від іншого концепту, що порівнюється, у визначенні міри розглядається мінімальне розширення концептів.

Визначення оцінок подібності на рівні знань та змішаному рівні

Нагадаємо, що до оцінок цих рівнів відносяться оцінки визначення ступеня подібності екземплярів та екземпляра і концепту. Визначення міри із залученням екземплярів базується на понятті Найбільш Специфічного Концепту (MSC). Для кожного екземпляра в ABox можна обчислити MSC або його апроксимацію. В деяких випадках ці поняття є еквівалентними.

Нехай *a* і *b* – два екземпляри ABox, *A*=MSC*(a), B*=MSC*(b)*. Тоді міри семантичної подібності можуть бути застосовані до описів концептів *A** та *B**, й результуюча оцінка виражатиме ступінь подібності відповідних екземплярів:

$$\forall a,b\colon s(a,b)=s(A^*,B^*)=s(MSC^*(a),MSC^*(b))$$

Аналогічно, значення подібності між описом концепту C та екземпляра a може бути обчислене шляхом визначення апроксимації MSC екземпляра та наступного застосування міри подібності до концепта C та апроксимації MSC* екземпляра a:

$$\forall a,C\colon s(a,C)=s(A^*,C)=s(MSC^*(a),C)$$

Тож, обидві оцінки зводяться до визначення подібності описів концептів після попередньої апроксимації екземплярів. Одночасно можуть бути використані будь-які наведені вище моделі обчислення ступеня відповідності концептів.

Слід зазначити, що складність запропонованих методів залежить від складності стандартних методів виведення в ДЛ.

Застосування оцінок подібності на прикладі ДЛ онтології POGeometry

Розглянемо застосування міри відповідності описів концептів на основі їхнього розширення у канонічній інтерпретації ДЛ на прикладі онтології домену геометричних понять *POGeometry*.

TBox онтології домену *POGeometry:*

Coordinate, GeometricFigure

Vertex ⊑ *has.XCoordinate*

Vertex ⊑ has.*YCoordinate*

XCoordinate ⊑ *Coordinate*

YCoordinate ⊑ *Coordinate*

Coordinate ⊑ *hasValue.NUMBER*

Vector ⊑ *2has.Vertex*

Vector ⊑ h*as.VectorLength*

Vector ⊑ *has.VectorAngle*

VerterxLength ⊑ *hasType.NUMBER*

VerterxAngle ⊑ *hasType.NUMBER*

Height ⊑ *hasType.NUMBER*

EdgeLenth ⊑ *hasType.NUMBER*

...

Polygon ≡ *GeometricFigure* ⊓=*has.Vertex*⊓ =*has.Vector*

Circle ⊑ *GeometricFigure*

Quadrangle ≡ *Polygon* ⊓=*4has.Vertex*⊓ =*4has.Vector*

Triangle ≡ *Polygon* ⊓ =*3has.Vertex* =*3has.Vector*

Polygon ⊑ *has.Vertex*

Polygon ⊑ *has.Vector*

Triangle ⊑ =*3has.Height*

Square ⊑ *hasType.NUMBER*

GeometricFigure ⊑ *has.Square*

Circle ⊑ *GeometricFigure*

ABox:

Triangle(ABC), Triangle(XYZ), Triangle(A1B1C1), Triangle(B1C1D1), Triangle(A1C1D1), Triangle(A1B1D1), Triangle(X1X2X3), Triangle(X2X3X4), Triangle(X3X4X5), Triangle(X4X5X6), ..., Quadrangle(A1B1C1D1), Polygon(X1X2X3X4X5X6.), Circle(O_1), Circle(O_2)

Враховуючи визначення концептів *Quadrangle* та *Triangle* можна вивести включення концептів *Triangle* ⊑ *Polygon* та *Quadrangle* ⊑ *Polygon*. Отже, всі екземпляри концептів *Triangle* та *Quadrangle* є екземплярами концепту *Polygon*.

Тобто $|Polygon^{\mathcal{I}}|$=47, $|Triangle^{\mathcal{I}}|$=29, $|Quadrangle^{\mathcal{I}}|$=17.

Тоді відповідність концептів *Triangle* та *Polygon* можна визначити на основі їхніх множин екземплярів таким способом:

Нехай $I = Triangle \sqcap Polygon,$, тоді

$$s(Polygon, Triangle) = \\ = \frac{|I^{\mathcal{I}}|}{|Polygon^{\mathcal{I}}| + |Triangle^{\mathcal{I}}| - |I^{\mathcal{I}}|} * \\ * \max\left(\frac{|I^{\mathcal{I}}|}{|Polygon^{\mathcal{I}}|}, \frac{|I^{\mathcal{I}}|}{|Triangle^{\mathcal{I}}|}\right) \\ = \frac{29}{47 + 29 - 29} * \max\left(\frac{29}{47}, \frac{29}{29}\right) = \frac{29}{47} = 0{,}62$$

Зважаючи на те, що інтерпретації концептів *Triangle* та *Quadrangle*, в даному випадку не мають перетину, $|I^{\mathcal{I}}| = 0$, де $I = Triangle \sqcap Quadrangle$, їх оцінка відповідності за екземплярами також буде дорівнювати 0. В цьому випадку, напевне, більш достовірнішими будуть оцінювання ступенів відповідності концептів за їхніми властивостями або з використанням найменшого спільного покриття.

Слід зазначити, що наведений приклад грунтується на базовій дескриптивній логиці, що використовує лише конструктор перетину, а TBox фактично є таксономією. Тому, для його концептів LCS існує завжди і для будь-яких концептів C і D з цього TBox *LCS(C,D)=MSA(C,D)*. Зокрема, $Polygon$ =*LCS*($Triangle$, $Quadrangle$) = MSA($Triangle$, $Quadrangle$).

Функцію подібності концептів на базі LCS можна визначати на основі відстаней між концептами або перекриття розширень відповідних концептів (їхніх множин екземплярів).

$$dist(Triangle, Quadrangle) =_{def} |nf(Quadrangle)| + \\ + |nf(Triangle)| - 2 * \left|nf\left(lcs(Triangle, Quadrangle)\right)\right| = \\ |nf(Quadrangle)| + |nf(Triangle)| - \\ - 2 * |nf(Polygon)|=2+2-2*1=2$$

Використовуючи модель властивостей, оцінка подібності:

$$s(Triangle, Quadrangle) = \\ =_{def} \frac{2f(ftrs(Triangle) \cap ftrs(Quadrangle))}{f\left(ftrs(Triangle)\right) + f\left(ftrs(Quadrangle)\right)} = \\ = \frac{2 * 1}{3 + 3} = \frac{1}{3}$$

Зважаючи на те, що в даному випадку GCS=LCS=MSA:

$$\mathcal{S}(Triangle, Quadrangle) \\ = \frac{\min\left(|Triangle^{\mathcal{I}}|, |Quadrangle^{\mathcal{I}}|\right)}{|(LCS(Triangle, Quadrangle))^{\mathcal{I}}|} \\ * \left(1 - \frac{\left|\left(LCS(Triangle, Quadrangle)\right)^{\mathcal{I}}\right|}{|\Delta^{\mathcal{I}}|} * \right. \\ \left. * \left(1 - \frac{\min(|Triangle^{\mathcal{I}}|, |Quadrangle^{\mathcal{I}}|)}{\left|\left(LCS(Triangle, Quadrangle)\right)^{\mathcal{I}}\right|}\right)\right) \\ = \frac{\min(|Triangle^{\mathcal{I}}|, |Quadrangle^{\mathcal{I}}|)}{|Polygon^{\mathcal{I}}|} \\ * \left(1 - \frac{|Polygon^{\mathcal{I}}|}{|\Delta^{\mathcal{I}}|} * \left(1 - \frac{\min(|Triangle^{\mathcal{I}}|, |Quadrangle^{\mathcal{I}}|)}{|Polygon^{\mathcal{I}}|}\right)\right) = \\ = \frac{17}{47} * \left(1 - \frac{47}{49} * \left(1 - \frac{17}{47}\right)\right) = \frac{17}{47} * \frac{19}{49} \approx 0{,}14$$

Висновки

У роботі здійснено аналіз семантичних показників подібності, які класифіковані за підходами та моделями оцінювання. Наведені показники використовують семантичні висновки, такі, як, наприклад, перевірка екземплярів (що означає обчислення розширення концептів) заданого ABox. Внутрішня складність виразних мов ДЛ, таких, як ALC, обумовлює неефективність структурних підходів до аналізу. Тому визначення функцій подібності базуються на використанні теорії множин, що дозволяє застосовувати числові підходи, хоча й на символічному рівні представлення ДЛ.

Проаналізовані також моделі оцінювання та міри подібності на різних рівнях оцінювання. Основним є встановлення подібності між концептами (моделі понятійного рівня). Задачі обчислення ступеня подібності між екземплярами та між концептом та екземпляром зводяться до знаходження MSC та оцінювання подібності концептів.

Більшість наведених оцінок грунтуються на основі базових ДЛ, що підтримують лише констуктор перетину, але описані підходи можуть бути застосовані для будь-якої ДЛ. Це забезпечує базові сервіси міркувань, а саме: перевірку екземплярів та MSC (апроксимацію).

Представлені міри подібності можуть бути корисними при вирішенні багатьох задач різних типів, зокрема задач великих даних, таких як-от, пошук інформації в контексті термінологічних систем представлення знань, категоризація та класифікація даних тощо.

Література

1. Fellbaum, C. (Ed.). (1998). Wordnet: An Electronic Lexical Database. MA: MIT Press.
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)
3. Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer (2004)
4. Thompson, K., Langley, P.: Concept formation in structured domains. In Fisher, D., Pazzani, M., Langley, P., eds.: Concept Formation: Knowledge and Experience in Unsupervised Learning. Morgan Kaufmann (1991)
5. Haussler, D.: Learning conjuntive concepts in structural domains. Machine Learning (1989) 7–40
6. F. Baader, R. K¨usters, and R. Molitor. Computing least common subsumers in description logics with existential restrictions. In T. Dean, editor, Proceedings of the 16th International Joint Conference on Artificial Intelligence, pages 96–101. Morgan Kaufmann, 1999.
7. F. Baader, R. Sertkaya, and Y. Turhan. Computing least common subsumers w.r.t. a background terminology. In V. Haarslev and R. M¨oller, editors, Proceedings of Proceedings of the 2004 International Workshop on Description Logics (DL2004). CEUR-WS. org, 2004.
8. R. Rada, H. Milli, E. Bicknell, M. Blettner, ”Development and Application of a metric on Semantic Nets”, IEEE Trans. on Systems, Man, and Cybernetics, 19(1): 17-30 (1989)
9. Mantay, T.: Commonality-based ABox retrieval. Technical Report FBI-HH-M-291/2000, Department of Computer Science, University of Hamburg, Germany (2000)
10. Collet, C., Huhns, M.N., Shen, W.M.: Resource integration using a large knowledge base in carnot. IEEE Computer 24 (1991) 55–62
11. Fankhauser, P., Neuhold, E.J.: Knowledge based integration of heterogeneous databases. In Hsiao, D.K., Neuhold, E.J., Sacks-Davis, R., eds.: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5). IFIP Transactions, North-Holland (1992)
12. Bright, M.W., Hurson, A.R., Pakzad, S.H.: Automated resolution of semantic heterogeneity in multidatabases. ACM Transaction on Database Systems 19 (1994) 212–253
13. Tversky, A.: Features od similarity. Psycological Review 84 (1997) 327–352
14. Jang, J., Conrath, D.: Semantic symilarity based on corpus statistic and lexical taxonomy. In: Proceedings of the International Conference on Computational Linguistics. (1997)
15. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11 (1999) 95–130
16. Weinstein, P., Bimingham, P.: Comparing concepts in differentiated ontologies. In: Proceedings of 12th Workshop on Knowledge Acquisition, Modelling, and Management. (1999)
17. Rodr´ıguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Transaction on Knowledge and Data Engineering 15 (2003) 442–456
18. A. Tversky, “Features of Similarity”, Psychological Review 84(4): 327-352, 1977.
19. J. Lee, M. Kim, and Y. Lee. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation, 2(49):188–207, 1993.

20. D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In Proceeding of the EON 2006 Workshop, 2006.
21. P. Resnik, ”Using Information Content to Evaluate Semantic Similarity”, Proc. IJCAI 1995 : 448-453
22. G. Miller & W.G. Charles, ”Contextual correlates of semantic similarity”, Language and Cognitive Processes, 6, 1-28, 1991.
23. W. Cohen, A. Borgida, H. Hirsh: “Computing Least Common Subsumers in Description Logics”, AAAI 1992: 754-760
24. R. Kusters & R. Molitor, “Computing Least Common Subsumers in ALEN”, IJCAI 2001: 219-224
25. Claudia d’Amato, Steffen Staab, Nicola Fanizzi, F. Esposito: “Efficient Discovery of Services Specified in Description Logics Languages”, SMRR 2007
26. C. d’Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In A. Pettorossi, editor, Proceedings of Convegno Italiano di Logica Computazionale, CILC05, Rome, Italy, 2005
27. C. d’Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for ALC concept descriptions. In Proc. of the 21st Annual ACM Symposium of Applied Computing, SAC2006, 2006.
28. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95–130, 1999.
29. A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, Proceedings of the 2005 International Workshop on Description Logics (DL2005), volume 147 of CEURWorkshop Proceedings. CEUR-WS. org, 2005.

References

1. Fellbaum, C. (Ed.). (1998). Wordnet: An Electronic Lexical Database. MA: MIT Press.
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)
3. Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer (2004)
4. Thompson, K., Langley, P.: Concept formation in structured domains. In Fisher, D., Pazzani, M., Langley, P., eds.: Concept Formation: Knowledge and Experience in Unsupervised Learning. Morgan Kaufmann (1991)
5. Haussler, D.: Learning conjuntive concepts in structural domains. Machine Learning (1989) 7–40
6. F. Baader, R. K¨usters, and R. Molitor. Computing least common subsumers in description logics with existential restrictions. In T. Dean, editor, Proceedings of the 16th International Joint Conference on Artificial Intelligence, pages 96–101. Morgan Kaufmann, 1999.
7. F. Baader, R. Sertkaya, and Y. Turhan. Computing least common subsumers w.r.t. a background terminology. In V. Haarslev and R. M¨oller, editors, Proceedings of Proceedings of the 2004 International Workshop on Description Logics (DL2004). CEUR-WS.org, 2004.
8. R. Rada, H. Milli, E. Bicknell, M. Blettner, ”Development and Application of a metric on Semantic Nets”, IEEE Trans. on Systems, Man, and Cybernetics, 19(1): 17-30 (1989)
9. Mantay, T.: Commonality-based ABox retrieval. Technical Report FBI-HH-M-291/2000, Department of Computer Science, University of Hamburg, Germany (2000)
10. Collet, C., Huhns, M.N., Shen, W.M.: Resource integration using a large knowledge base in carnot. IEEE Computer 24 (1991) 55–62
11. Fankhauser, P., Neuhold, E.J.: Knowledge based integration of heterogeneous databases. In Hsiao, D.K., Neuhold, E.J., Sacks-Davis, R., eds.: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5). IFIP Transactions, North-Holland (1992)
12. Bright, M.W., Hurson, A.R., Pakzad, S.H.: Automated resolution of semantic heterogeneity in multidatabases. ACM Transaction on Database Systems 19 (1994) 212–253
13. Tversky, A.: Features od similarity. Psycological Review 84 (1997) 327–352
14. Jang, J., Conrath, D.: Semantic symilarity based on corpus statistic and lexical taxonomy. In: Proceedings of the International Conference on Computational Linguistics. (1997)
15. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its ap-

plication to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11 (1999) 95–130
16. Weinstein, P., Bimingham, P.: Comparing concepts in differentiated ontologies. In: Proceedings of 12th Workshop on Knowledge Acquisition, Modelling, and Management. (1999)
17. Rodr´ıguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Transaction on Knowledge and Data Engineering 15 (2003) 442–456
18. A. Tversky, “Features of Similarity”, Psychological Review 84(4): 327-352, 1977.
19. J. Lee, M. Kim, and Y. Lee. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation, 2(49):188–207, 1993.
20. D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In Proceeding of the EON 2006 Workshop, 2006.
21. P. Resnik, ”Using Information Content to Evaluate Semantic Similarity”, Proc. IJCAI 1995 : 448-453
22. G. Miller & W.G. Charles, ”Contextual correlates of semantic similarity”, Language and Cognitive Processes, 6, 1-28, 1991.
23. W. Cohen, A. Borgida, H. Hirsh: “Computing Least Common Subsumers in Description Logics”, AAAI 1992: 754-760
24. R. Kusters & R. Molitor, “Computing Least Common Subsumers in ALEN”, IJCAI 2001: 219-224
25. Claudia d’Amato, Steffen Staab, Nicola Fanizzi, F. Esposito: “Efficient Discovery of Services Specified in Description Logics Languages”, SMRR 2007
26. C. d’Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In A. Pettorossi, editor, Proceedings of Convegno Italiano di Logica Computazionale, CILC05, Rome, Italy, 2005
27. C. d’Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for ALC concept descriptions. In Proc. of the 21st Annual ACM Symposium of Applied Computing, SAC2006, 2006.
28. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95–130, 1999.
29. A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, Proceedings of the 2005 International Workshop on Description Logics (DL2005), volume 147 of CEURWorkshop Proceedings. CEUR-WS. org, 2005.

Про автора:

Захарова Ольга Вікторівна,
кандидат технічних наук,
старший науковий співробітник.
Кількість наукових публікацій в українських виданнях – 31.
http://orcid.org/0000-0002-9579-2973.

Одержано: 02.04.2021

Місце роботи автора:

Інститут програмних систем НАН України,
проспект Академіка Глушкова, 40.
Тел.: 526 5139.
E-mail: ozakharova68@gmail.com.
Моб.тел.: +38(068)594756