УДК 681.324

В. М. Дубовой, д-р. техн. наук, проф.; О. М. Москвін, асп.

ОЦІНКА ГІПЕРТЕКСТОВОЇ СТРУКТУРИ НА ОСНОВІ СЕМАНТИЧНОЇ ІНФОРМАЦІЇ

Розглянуто проблему використання семантичної інформації для оцінки зв'язків гіпертекстової структури та алгоритм розв'язання цієї задачі з метою подальшої оптимізації цієї структури за швидкістю досяжності користувачем її фрагментів.

Велика кількість Інтернет-ресурсів, їх відносно низька якість, що виражена в хаотичній, неоптимізованій структурі, сповільнюють пошук необхідної інформації, яка зазвичай подана у вигляді гіпертексту. Отже, має місце проблема низької швидкості досяжності фрагментів гіпертексту у мережі ресурсів. Для вирішення проблеми в теорії гіпертексту введена спеціальна гіпертекстова метрика [1], недоліками якої є ненормованість показників та відсутність формальної моделі для оцінювання семантичної структури гіпертексту. Хоча авторами запропоновано нормовану метрику на основі нечіткої моделі [2], що оперує ієрархічними зв'язками, без урахування залишаються семантичні особливості ресурсів, які аналізуються.

На цей момент підтримка семантичної інформації набуває застосування у пошукових системах, наприклад, Yandex, але визначення важливості зв'язків є проблематичним. Для впливу на процес навігації користувача використовується аналіз структурних зв'язків без урахування їх семантики. До таких реалізацій можна віднести загальновідомий Google Page Rank.

Ідеї застосування та впровадження засобів машинного аналізу даних були запропоновані консорціумом W3 у 1999 році [3] і активно впроваджуються на практиці. Основною метою використання семантичної інформації є відхід від аналізу гіпертексту, призначеного для користувача, до формалізованих даних, призначених для машин та інтелектуальних агентів.

Метою роботи є розв'язання задачі аналізу і оцінювання гіпертекстової структури на основі семантичної інформації, що спрямовано на підвищення ефективності використання веб-ресурсів.

Пропонується проводити її аналіз у підмережі Інтернет-ресурсів для оцінювання гіпертекстових зв'язків з метою подальшої оптимізації досяжності в гіпертекстовій мережі. Причому очікується, що ресурси, які розглядаються, та, відповідно, їх інформаційні статті складають певну ієрархію та характеризуються наявністю онтологічної інформації, формалізованої за допомогою мов RDF чи OWL. Пропонована система оцінювання націлена на функціонування в умовах неповної інформації про загальну семантичну структуру гіпертексту, наявними є лише дані суміжних ресурсів, що є досить важливою умовою у разі функціонування у такому динамічному середовищі, як мережа Інтернет. Для ефективного семантичного аналізу необхідною умовою є здійснення процедури узгодження, тобто приведення до однієї ієрархії понять, онтологій Інтернет ресурсів, що розглядаються.

Узгодження онтологічних контекстів інформаційних статей, навіть за умови їх належності до схожих предметних областей, є проблематичним, оскільки має місце невизначеність понять. З цією метою пропонується застосувати низку засобів з їх попередньої інтеграції. Задача узгодження нетривіальна і, в основному, полуавтоматична, оскільки на цьому етапі розвитку онтологічних визначень тільки експерт може кінцево підтвердити коректність семантики встановлених відношень між онтологічними поняттями.

Для розв'язання задачі ідентифікації семантично пов'язаних понять необхідні методи, що дозволяють виділити групи схожих інформаційних статей і встановлювати зв'язки між множинами цих груп. Для розв'язання вказаної задачі виправданим є використання методів бі-кластеризації (об'єктно-ознакової кластеризації), в яких схожість об'єктів, що

об'єднуються у кластер, виражена через елементи опису всіх об'єктів цього кластеру.

До таких методів відносяться методи аналізу даних, основані на формальних поняттях і решітках формальних понять [4]. Апарат формального аналізу понять ФАП дозволяє формалізувати екстенсіонал і інтенсіонал об'єктів та їх взаємні відношення. Теорія решіток, що лежить в основі ФАП, дозволяє проводити класифікацію понять (концептів) з набору вхідних даних.

Використаємо необхідні визначення з [4]. Контекстом в ФАП є трійка $K := (G, M, I)$, де G — множина об'єктів; M — множина ознак (атрибутів); I — бінарне відношення між G та M, тобто $I \subseteq G \times M$. Для довільних $A \subseteq G$ і $B \subseteq M$ визначені оператори Галуа

$$\begin{aligned} A' &= \left\{ m \subseteq M \middle| \forall g \subseteq A : (g,m) \in I \right\}; \\ B' &= \left\{ g \subseteq G \middle| \forall m \subseteq B : (g,m) \in I \right\}, \end{aligned} \tag{1}$$

оператор $(\cdot)''$ (композиція двох застосувань оператора $(\cdot)'$) є операцією замикання — він є ідемпотентним $(A'' = A)$, монотонним $(A \subseteq B \Rightarrow A'' \subseteq B)$ і екстенсивним $(A \subseteq A'')$.

Формальне поняття контексту (G, M, I) — пара (A, B), така що $A \subseteq G$, $B \subseteq M$, $A = B'$ та $B' = A$, $A = A''$, $B = B''$. Множина A є екстенсіоналом (об'ємністю поняття), B — інтенсіоналом (змістовністю поняття).

Поняття (A_1, B_1) і (A_2, B_2) пов'язані відношенням часткового порядку $(A_1, B_1) \le (A_2, B_2)$, якщо $A_1 \subseteq A_2 (B_1 \supseteq B_2)$. В цьому випадку (A_1, B_1) називають менш загальним поняттям (A_2, B_2), а (A_2, B_2) — узагальненням поняття (A_1, B_1).

Частково впорядкована за вкладеністю об'ємів множина формальних понять контексту K позначається $\mathfrak{R}(K)$ і називається решіткою понять контексту K.

В задачах пошуку перетину онтологій і ідентифікації груп інформаційних статей як об'єктів пропонується використовувати унікальні ідентифікатори інформаційних статей, наприклад, URL, а в якості ознак — ключові поняття заданих на них онтологій. В цьому випадку діаграма решітки, що утворюється, є наочним образом таксономії груп (замкнутих множин) статей.

Оскільки загальна кількість понять, що породжується в багатьох випадках, є надлишковою і може заважати її аналізу через інформаційний «шум», важливою умовою є оцінка і фільтрація результату кластеризації для виділення «найважливіших» груп інформаційних статей. Пропонується використовувати індекс інтенсіональної стійкості [4], що показує, наскільки зміст формального поняття залежить від окремого об'єкта:

$$\sigma_i(A, B) = \frac{\left| \left\{ C \subseteq A \middle| C' = B \right\} \right|}{2^{|A|}}, \tag{2}$$

що фактично дорівнює відношенню потужності множини об'єктів C з інтенсіоналом B до кількості можливих підмножин об'єктів множини A, тобто індекс дорівнює частці підмножин множини об'єктів, що породжують це формальне поняття. Оскільки в цій роботі нас в першу чергу цікавить зміст понять, тобто ключові слова, що становлять онтології, будемо використовувати інтенсіональну стійкість і відбирати поняття з найстійкішими множинами онтологічних елементів.

Пропонується використання цього показника для побудови розподілу, що характеризує вірогідність перетину онтологічних понять, враховуючи, що область визначення $\sigma_i(A, B) = [0;1]$.

Незважаючи на значні переваги, що надають онтології для автоматизованого аналізу інформації, поданої в мережі Інтернет, з їх використанням пов'язана низка недоліків:

— низька швидкість обробки великих онтологій;

— відносно невелика розповсюдженість онтологічної інформації;

— неможливість створення єдиної узгодженої онтології у зв'язку із її складністю, частотою зміни та високою вартістю обслуговування;

— невизначеність понять;

— використання різноманітних мов та засобів опису онтологій.

Беручи до уваги вищенаведені недоліки, пропонується розглядати лише множину взаємопов'язаних ресурсів для аналізу з подальшою оптимізацією. Цей підхід дозволяє вирішити проблеми узгодження онтологій на глобальному рівні за рахунок розгляду ресурсів, що належать до одних і тих самих предметних областей, отже вірогідність невірного узгодження понять значно знижується. Крім того знаходить вирішення проблема, пов'язана з низькою швидкістю обробки онтологій — у випадку розгляду лише суміжних ресурсів до заданого, розмір узгодженої онтології значно зменшується. Обслуговування невеликих онтологій для ресурсів є відносно дешевою операцією як з точки зору складності, так і з точки зору використання людських ресурсів на її створення і обслуговування, порівняно з створенням глобальних таксономій. Тому використання розподіленого зберігання і обслуговування онтологій дозволяє значно ефективніше розв'язувати задачі керування семантичною мережею.

В умовах частої зміни інформації і, відповідно, її онтологічної складової, пропонується використання автоматизованих інтелектуальних агентів для розв'язання поставлених задач, для швидкої адаптації до навколишнього середовища, характеризованого частою зміною стану, та реалізації розподіленої його обробки, що дозволить значно підвищити операційну швидкість зазначених задач.

Запропонований підхід передбачає наявність інтелектуального програмного агента, прикріпленого до Інтернет-ресурсу, який аналізує гіпертекстові зв'язки з сусідніми ресурсами. Програми-агенти розміщуються на web-серверах. В процесі роботи агенти обмінюються інформацією щодо структури та онтологічної інформації ресурсів, до яких вони належать. В цьому випадку також передбачається, що кожна сторінка містить онтологічний контекст. В результаті аналізу, сутність якого наведена нижче, відбувається оцінка важливості існуючих зв'язків, яка у вигляді рекомендаційної інформації надається адміністратору ресурсу для подальшої їх зміни.

Для оперування над онтологіями та їх складовими пропонується використання багатоосновної алгебри, раніше запропонованої авторами. Алгебра включає операції об'єднання, перетину, віднімання, морфологічного узгодження, визначення ієрархій понять та визначення належності до них.

Оскільки можливі випадки різнорідних онтологій, що можуть мати морфологічно однакові, але семантично різні поняття, які автоматизована система оцінки може трактувати як ідентичні, то важливою задачею в процесі узгодження онтологій є оцінка рівня невизначеності об'єднаної таксономії для прийняття рішення про використання результату об'єднання для аналізу. Некоректне з точки зору семантики об'єднання онтологій може призвести до суперечливих результатів як оцінки, так і подальшої оптимізації, тому важливим є відшукання і уникнення агрегування онтологій в даному випадку.

Для оцінки невизначеності результату узгодження пропонується використання ентропії та її властивості ієрархічної адитивності [5].

Для цього побудуємо дерево вибору, що зображено на рис. 1, на основі 5 випадкових величин, зазначених в табл., які розподілені за рівномірним законом розподілу і приймають значення 0,5. Чим більше гілок пройдено під час аналізу результату об'єднання, тим більша ентропія і, відповідно, тим ненадійнішим є результат об'єднання.

Величини для оцінки невизначеності результату узгодження

Величина	Зміст	Значення
ε_1	Правила об'єднання онтологій відсутні	1 — ні, 2 — так
ε_2	Ієрархії понять мають різні корені	1 — ні, 2 — так
ε_3	Немає загальних типів	1 — так, 2 — ні
ε_4	Немає загальних супертипів	1 — ні, 2 — так
ε_5	Немає збігу після проведення морфологічного аналізу	1 — ні, 2 — так

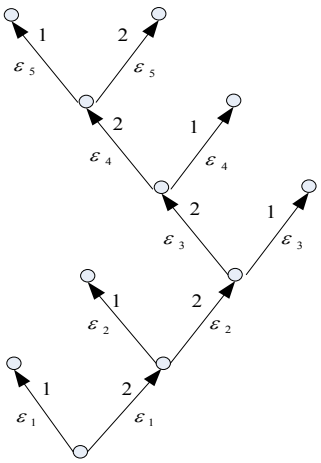

Рис. 1. Дерево вибору для невизначеності результату узгодження

Величини, обрані на основі експертного аналізу, дозволяють оцінити на базовому рівні, наскільки набір онтологій є взаємопов'язаним.

Згідно з [6]

$$H_{\varepsilon_1 \ldots \varepsilon_{n-1}} = H_{\varepsilon_2|\varepsilon_1} + H_{\varepsilon_2|\varepsilon_1\varepsilon_2} + H_{\varepsilon_2|\varepsilon_1\varepsilon_2\varepsilon_3} + H_{\varepsilon_2|\varepsilon_1 \ldots \varepsilon_{n-1}} \quad (3)$$

— ентропія вибору у дереві вибору, де M — кількість етапів розбиття; H_{ε_k} — ентропія вибору на k-му етапі у вузлі, а повна ентропія k-го кроку у загальному вигляді

$$H_{\varepsilon_k|\varepsilon_1,\ldots,\varepsilon_{k-1}} = MH_{\varepsilon_k}\left(|\varepsilon_1,\ldots,\varepsilon_{k-1}\right) = MP\left(\varepsilon_k|\varepsilon_1,\ldots,\varepsilon_{k-1}\right)\cdot \ln P\left(\varepsilon_k|\varepsilon_1,\ldots,\varepsilon_{k-1}\right). \quad (4)$$

Узагальнена схема методу зображена на рис. 2.

Як зазначено раніше, будемо розглядати онтології з однорідними та семантично пов'язаними таксономіями для розв'язання задачі пошуку глобальної відповідності.

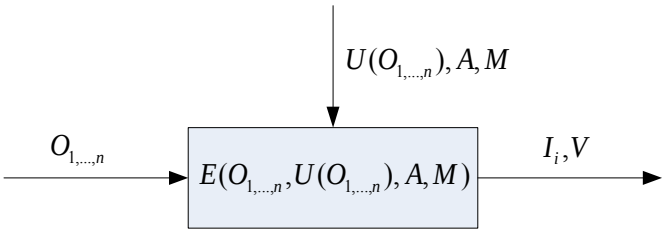

Рис. 2. Загальна схема методу пошуку перетину онтологій

На вхід методу подається онтологія інформаційної статті O_i, для якої необхідно знайти рівень зв'язності з сусідніми інформаційними статтями. Дані про всі суміжні інформаційні одиниці та їх онтології подані $U(O)$, що надається системі оцінки автоматизованим агентом. Для здійснення семантично коректної ідентифікації понять метод приймає на вхід правила встановлення відповідності між онтологіями A і вектор словоформ M, який надається із зовнішнього джерела даних, наприклад, *www.morphology.ru*, для здійснення уточнення шляхом морфологічного аналізу.

Результатом здійснення процедури пошуку перетину є пара векторів (I_i, V_i), де I_i — значущі поняття з рівнем інтенсіональної стійкості не нижче заданої, а V_i — значення стабільності для поняття I_i. Порогове значення стабільності σ_k для відбору сутностей задається на основі експертного аналізу в інтервалі $(0,5; 1]$.

Запропонований метод складається з таких етапів:

1. Визначення автоматизованим агентом $U(O)$.

2. Визначення ентропії об'єднаної онтології $H_{\varepsilon_1 \ldots \varepsilon_{n-1}}$.

3. У випадку, коли $H_{\varepsilon_1 \ldots \varepsilon_{n-1}} > 0,5$, онтології вважаються несумісними і в результаті не піддаються аналізу на перетин екстенсіоналів.

4. Визначення вектора A та перетворення правил у відношення $\xi(O_1, O_2, A_i)$ поточної інформаційної статті — для кожної онтології з $U(O)$.

5. Проведення операції узгодження $\omega(O_1, O_2, A_1, A_2)$ з A.

6. Проведення морфологічного узгодження $\omega_M(O_1, O_M, M)$, визначеного в [5].

7. Побудова решітки формальних понять для множини інформаційних статей з суміжних ресурсів.

8. Обчислення індексу інтенсіональної стабільності $\sigma_i(A,B)$ для екстенсіоналів, що розглядаються.

9. Формування рекомендацій обслуговуючому персоналу ресурсу щодо необхідності вилучення або додавання зв'язків між ресурсами на основі значення індексу інтенсіональної стабільності σ_i. Якщо $\sigma_i < \sigma_k$, то рекомендацією буде видалення відповідного зв'язку, і навпаки, для $\sigma_i \geq \sigma_k$ — його додавання.

Отже, запропонована схема пошуку перетинів онтологій і метод оцінки важливості зв'язків між інформаційними одиницями, в якому на відміну від існуючих підходів, застосовують семантичну інформацію, що спеціально призначена для інтелектуального аналізу гіпертекстових даних та формальний аналіз понять, в основі якого лежить математичний апарат теорії решіток.

Оскільки задача оцінювання зводиться до обчислень на матрицях, її розв'язання можна пришвидшити за допомогою розпаралелювання обчислювального процесу.

Висновки

Запропонований підхід до оцінки зв'язків підмережі інформаційних ресурсів для її подальшої оптимізації на основі семантичної інформації забезпечить оптимальну досяжність фрагментів гіпертексту в умовах неповної інформації про структуру мережі.

Запропонований підхід до оцінки результату узгодження онтологій забезпечує семантичну коректність контекстів.

Отже, комплексне використання запропонованої алгебри для оперування над онтологічними даними, методу оцінки точності узгодження онтологічних контекстів та методу оцінки онтологічної інформації сприятиме підвищенню ефективності використання Інтернет-ресурсів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Botafogo R. A. Identifying hierarchies and useful metrics / R. A. Botafogo, Rivlin E., Shneiderman B. // ACM Transactions on Information Systems (TOIS). — 1992. — No 2. — P. 142—180.

2. Розробка нечіткої системи класифікації гіпертекстових структур [Електронний ресурс] / В. М. Дубовой, О. М. Москвін // Наукові праці ВНТУ. — 2009. — № 2. — Режим доступу : http://www.nbuv.gov.ua/e-journals/vntu/2009-2/2009-2.files/uk/09vmdfcs_ua.pdf.

3. The Semantic Web [Електронний ресурс] / Tim Berners-Lee, James Hendler, Ora Lassila // Scientific American Magazine. — May, 2001. — Режим доступу до журналу : http://www.sciam.com/article.cfm?id=00048144-10D2-1C70-84A9809EC588EF21.

4. Towards Concise Representation for Taxonomies of Epistemic Communities [Електронний ресурс] / Camille Roth, Sergei Obiedkov, Derrick G. Kourie // CLA 4th Intl Conf on Concept Lattices and their Applications, Tunis, Tunisia. — Oct, 2006. — Режим доступу : http://www.springerlink.com/content/a822lw11w1ugn74n/.

5. Стретович Р. Л. Теория информации / Р. Л. Стретович. — М. : Сов. Радио, 1975. — 424 с.

Рекомендована кафедрою комп'ютерних систем управління

Стаття надійшла до редакції 25.02.11
Рекомендована до друку 11.03.11

Дубовой Володимир Михайлович — завідувач кафедри, ***Москвін Олексій Михайлович*** — аспірант.

Кафедра комп'ютерних систем управління, Вінницький національний технічний університет, Вінниця