УДК 004.896

Автоматизована побудова дидактичної онтології на основі Wikipedia

Левченко Я.А., Титенко С.В., к.т.н.
Національний технічний університет України «КПІ», м. Київ
yaroslav.levchenko1@gmail.com

Левченко Я. А. Автоматизована побудова дидактичної онтології на основі Wikipedia / Я. А. Левченко, С. В. Титенко // Международная научная конференция имени Т.А. Таран «Интеллектуальный анализ информации» ИАИ-2015, Киев, 20–22 мая 2015 г. : сб. тр. – К. : Просвіта, 2015. – 131-137 с.

У роботі запропонований формальний апарат визначення дидактичного порядку статей Wikipedia, що ґрунтується на нечіткій логіці Б'юкенена. Були використані алгоритми Бергера-Шора та топологічного сортування графу. Наведені результати роботи програми, в якій реалізовано запропонований формальний апарат.

Вступ

Wikipedia – вільна загальнодоступна багатомовна універсальна інтернет-енциклопедія, реалізована на принципах вікі. Вона містить понад 30 мільйонів статей. Інтернет-сайт Вікіпедії є п'ятим за відвідуваністю сайтом у світі. За обсягом відомостей і тематичним охопленням вона вважається найповнішою енциклопедією з коли-небудь створених за всю історію людства. Вікіпедія неухильно набувала статус загального довідкового сайту з моменту її створення.

Величезна кількість інформації, що розміщена на Вікіпедії, може слугувати матеріалом для самонавчання, але відбір та впорядкування цієї інформації потребують значних зусиль. Отже актуальною задачею є подання ресурсів Вікіпедії у зручній послідовності з урахуванням дидактичних зв’язків, що суттєво скоротило б час на пошук потрібної навчальної інформації.

Метою цієї роботи є дидактичне впорядкування інформації відповідно до запиту користувача. Вхідними даними слугує посилання на статтю. За вхідними даними завдяки автоматизованому аналізу Вікіпедії формується дидактична онтологія. На основі онтології на вихід

подається впорядкована вибірка статей, що дидактично зв'язані з вхідним поняттям.

Вибірка цільової множини зв'язаних понять

Система отримує на вхід від користувача запит на вивчення певного поняття, поданого статтею Вікіпедії. Нехай c_0 – цільова стаття. Наступним завданням системи є пошук такої множини понять, які потенційно можуть брати участь у результуючому наборі матеріалів для вивчення. Такі поняття визначаються на основі аналізу перехресних *посилань* між статтями. Так як для статей Вікіпедії характерна висока насиченість посиланнями на статті з інших предметних областей, що не є значущими при цілеспрямованому предметному вивченні деякої теми, виникає потреба відсіювання зайвих посилань. Нижче пропонуються етапи отримання множини понять, що пройшли процес тематичного відбору.

Множина посилань $Link = \{c: (c_i, c_k)\}$[1] – множина пар таких понять, що існує, посилання зі статті c_i в статтю c_k.

Множина зв'язаних посилань A формується на основі двох рівнів вкладенності, тобто вибираються всі посилання з цільової статті, потім з кожного з цих посилань також вибираються всі посилання.

$Leve1(c_0) = \{ c: (c_0, c) \in Link \}$[1] – множина понять, що належать першому рівню вкладеності статей.

$Leve2(c_0) = \{ c: c_k \in Leve1(c_0) \wedge (c_k, c) \in Link \}$[1] – другий рівень вкладеності статей.

$A = \{ c: c_k \in Leve1(c_0) \cup Leve2(c_0) \wedge (c_k, c) \in Link \}$.

Наступним кроком є формування множини посилань B, що входять до всіх категорій цільового посилання. Результуючою множиною понять R буде перетин множин A та B: $R = \{ c: c_k \in A \cap B \wedge (c_k, c) \in Link \}$. Саме над результуючою множиною будуть проводитись усі наступні операції по визначенню дидактичних відношень між поняттями.

Правила дидактичного впорядкування

Усі посилання із множини R формують вершини графа онтології G. Ребра між вершинами графа формуються автоматично на основі аналізу тексту статей з використанням нечіткої логіки Б'юкенена [2] на основі прикладу використання даного апарату для дидактичних задач, запропонованих в роботі [3]. Логіка визначення дидактичного порядку

ґрунтується на правилах, кожному з яких відповідає певний фактор впевненості CF.

Далі подано правила визначення зв'язку між двома поняттями c_1 та c_2 із множини відібраних понять *R*.

Правило 1. Якщо в тексті статті c_2 буде знайдено посилання на статтю c_1, то це буде означати, що стаття c_1 дидактично передує статті c_2 з фактором впевненості, що залежить від віддаленості посилання від початку статті:

$$(c_2, c_1) \in Link \rightarrow concept_{before(c_1,c_2)} < CF = 0{,}3 * \left(1 - \frac{Dist}{Leng}\right) >,$$

де $Dist$ – віддаленість посилання на статтю c_1 в тексті статті c_2 в символах, $Leng$ – довжина тексту статті c_2 в символах.

Посилань в статті c_2 на статтю c_1 може бути декілька (n). Результуючий фактор впевненості даного правила для усіх *n* посилань визначається послідовним додаванням фактору впевненості CF до результуючого фактору CF_{end} відповідно до формул Б'юкенона [2]:

$$CF_1 + CF_2 = CF_1 + CF_2 - CF_1 * CF_2.$$

Правило 2. Якщо в першому абзаці статті c_2 було знайдено посилання на статтю c_1, то це означає, що стаття c_1 дидактично передує статті c_2:

$$(c_2, c_1) \in FirstIndent \rightarrow concept\_before(c_1, c_2) < CF = 0{,}4 >,$$

де $FirstIndent = \{c : (c_i, c_k) \in Link\}$ – множина пар таких понять, що в першому абзаці статті c_i існує посилання на статтю c_k.

Правило 3. Входження назви поняття c_1 в назву поняття c_2. Це свідчить, що стаття c_1 дидактично передує статті c_2:

$$(c_2, c_1) \in CInC \rightarrow concept\_before(c_1, c_2) < CF = 0{,}9 >,$$

де $CInC = \{c : (c_i, c_k) \in Link\}$ – множина пар таких понять, що в назві статті c_i знайдена назва статті c_k.

Правило 4. Якщо було знайдено входження wiki-посилання на поняття c_1 у тексті wiki-посилання на поняття c_2, то це буде показувати, що стаття c_1 дидактично передує статті c_2:

$$(c_2, c_1) \in TInT \rightarrow concept\_before(c_1, c_2) < CF = 0{,}9 >,$$

де $TInT = \{c : (c_i, c_k) \in Link\}$ – множина пар таких понять, що в wiki-посиланні статті c_i знайдено wiki-посилання статті c_k.

Таким чином, на те, що стаття c_1 дидактично передує статті c_2, може вказувати декілька факторів впевненості (n). Результуючим фактором

CFr буде результат послідовного додання всіх факторів . Додавання факторів впевненості проводиться як у випадку з першим правилом.

Усунення конфліктів у відношеннях дидактичного порядку

При аналізі дидактичного відношення між двома поняттями можлива ситуація, коли одночасно існують свідчення на користь протилежних гіпотез дидактичного порядку. Формально така ситуація виражається у наявності двох протилежних ребер між двома вершинами графа онтології. Таким чином, існує фактор впевненості, що показує достовірність того, що стаття c_1 дидактично передує статті c_2 — $concept\_before(c_1, c_2) < CF_{(c_1 \to c_2)} >$, і також існує фактор впевненості у тому, що стаття c_2 дидактично передує статті c_1 — $concept\_before(c_2, c_1) < CF_{(c_2 \to c_1)} >$. У такому випадку достовірним вважається відношення з більшим *CF*, а сам фактор впевненості перераховується за формулою [2, 3]:

$$CF = \frac{MAX\left(CF_{(c_1 \to c_2)}, CF_{(c_2 \to c_1)}\right) - MIN\left(CF_{(c_1 \to c_2)}, CF_{(c_2 \to c_1)}\right)}{1 - MIN\left(CF_{(c_1 \to c_2)}, CF_{(c_2 \to c_1)}\right)}$$

Граф дидактичної онтології та результуюча послідовність понять

Для вибірки R розраховуються всі фактори впевненості та усуваються конфлікти. На основі множини факторів впевненості формується матриця суміжності графу G.

Граф дидактичної онтології вказує на порядок вивчення понять, і, таким чином не повинен містити циклів. Натомість у результаті автоматичного аналізу Вікіпедії та застосування логічного апарату отриманий граф *G* може містити цикли, які підлягають подальшому усуненню. Для виконання даного завдання було обрано алгоритм Бергера-Шора [4]. Ациклічний орграф дидактичної онтології перетворюється у лінійну послідовність навчальних понять за допомогою алгоритму топологічного сортування графу. Таким чином, вибірка статей R подається користувачу у вигляді лінійної дидактичної послідовності як рекомендаційний інформаційно-навчальний перелік статей для детального вивчення цільового поняття.

Етапи роботи програмного комплексу

Вхід. Користувач подає на вхід посилання на статтю, яка його цікавить. Наприклад:

https://ru.wikipedia.org/wiki/Декларативное_программирование.

1. Синтаксичний аналіз Вікіпедії (парсинг)

1) GET-запит до API Wikipedia, який дає весь контент цільової статті;

2) Парсинг контенту, а саме витяг усіх посилань, самої статті, категорій.

2. Формування вибірки статей

1) Із цільової статті беруться всі посилання, для яких повторюється пункт 2;

2) Формується початкова множина A, яка складається з цільової статті, посилань з цієї статті та посилань з посилань цільової статті;

3) Формується множина B, що складається зі статей, що входять до всіх категорій цільової статті;

4) Формується кінцева множина статей R, що є перетином двох попередніх множин: $R = A \cap B$.

3. Розрахунок факторів впевненості

1) Для множини статей R обчислюються всі можливі фактори впевненості $concept\_before(c_i, c_k) < CF >$;

2) Вибираються всі фактори, що більше нуля;

3) Усуваються конфлікти.

4. Формування матриці суміжності на основі факторів впевненості

1) Формування матриці суміжності графу онтології G;

2) Застосування до графу G формули Флойда-Варшала;

3) Розрив циклів в графі G за методом Бергера-Шора;

4) Топологічне сортування графу G.

5. Виведення результатів.

1) Візуалізація графу онтології G;

2) Виведення результатів топологічного сортування графу онтології G.

Приклад роботи програмного комплексу

Для прикладу за цільове поняття було вибрано «Объектно-ориентированное программирование» (рос.). Вибірка посилань спеціально зменшена для наочності та складається з понять:

1) 'Объектно-ориентированное программирование';
2) 'Класс (программирование)';
3) 'Метод (программирование)';

4) 'Объект (программирование)';
5) 'Инкапсуляция (программирование)';
6) 'Наследование (программирование)';
7) 'Полиморфизм (информатика)';
8) 'Свойство (программирование)';

Результат топологічного сортування:
«Читайте перед статьей «Объектно-ориентированное программирование»:
1) Класс (программирование)
2) Метод (программирование)
3) Объект (программирование)
Читайте дополнительно:
1) Наследование (программирование)
2) Полиморфизм (информатика)
3) Инкапсуляция (программирование)».

Візуалізація графу онтології.

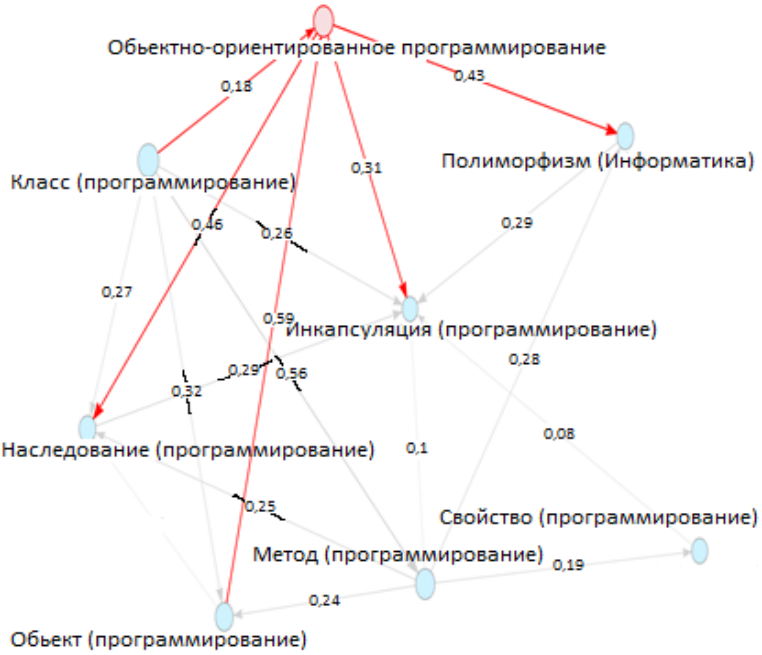

Рис.1. Візуалізація графу онтології

Висновок

У роботі проведено аналіз особливостей структури статей Вікіпедії та перехресних посилань щодо визначення дидактичних відношень між поняттями онлайн-енциклопедії. На основі аналізу запропоновано набір правил нечіткого виведення для визначення дидактичного порядку статей Вікіпедії, а також додаткова обробка результату роботи апарату нечіткого виведення. Результат у вигляді графу дидактичної онтології на базі відношень дидактичного порядку дозволяє будувати адекватні послідовності вивчення понять цільової предметної області, що спрощує процес самонавчання. Розроблений програмний комплекс реалізує запропонований формальний апарат.

Перспективними напрямками подальших досліджень є виявлення додаткових закономірностей в структурі статей Вікіпедії, що можуть бути використані для дидактичного аналізу, розробка евристичних алгоритмів розриву циклів, які б враховували наявні дані про достовірність дидактичних відношень у графі онтології. Додаткової уваги потребує проблема неповноти або некоректності оформлення текстів та перехресних посилань Вікіпедії, що часом має місце та призводить до некоректних результатів. Поточні дослідження представлені на сайті www.setlab.net.

Література

1. Супряга І. А. Система автоматизованої побудови навчальних ресурсів на основі статей Wikipedia/ І. А.Супряга, С. В. Титенко// First international forum «IT-Trends: big data, artificial intelligence, social media»:Book of abstracts. – Kremenchuk: Kremenchuk Mykhailo Ostrohradskiy National University, 2014. – С. 49-51.
2. Buchanan B. G., Shortliffe E. H. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. – MA: Addison-Wesley, 1984. – 769 p.
3. Титенко, С. В. Побудова дидактичної онтології на основі аналізу елементів понятійно-тезисної моделі/ С. В. Титенко // Наукові вісті НТУУ "КПІ". – 2010. – № 1(69). – С. 82-87.
4. Berger B., Shor P. W. Approximation alogorithms for the maximum acyclic subgraph problem //Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms. – Society for Industrial and Applied Mathematics, 1990. – С. 236-243.