

СИСТЕМА АВТОМАТИЗОВАНОЇ ПОБУДОВИ НАВЧАЛЬНИХ РЕСУРСІВ НА ОСНОВІ СТАТЕЙ WIKIPEDIA

І. А.Супряга, С. В. Титенко

Національний технічний університет України «Київський політехнічний інститут»
проспект Перемоги, 37, м.Київ, 03056. E-mail: supryaga.igor@gmail.com

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.

Метою даної роботи є дослідження та розробка методів автоматичної побудови інформаційно-навчальних ресурсів на базі Wikipedia відповідно до навчальної мети користувача. Було проведено аналіз структури статей Wikipedia та зв'язків між ними. Розроблено логічні правила визначення тематичної спорідненості понять, що ґрунтуються на використанні трьох рівнів вкладеності. Запропоновано правила визначення дидактичного порядку між статтями Wikipedia. На основі аналізу особливостей структури статей Wikipedia було розроблено правило визначення дидактичного порядку статей за позицією входження. Успішно застосовано правило визначення дидактичного порядку за іменем до статей Wikipedia

Для вирішення задачі необхідно реалізувати наступні завдання:

- розглянути аналогічні системи;
- розробити метод визначення споріднених понять відповідно до деякого поняття Wikipedia;
- розробити метод визначення дидактичного порядку між статтями Wikipedia;
- на основі запропонованих методів реалізувати дослідний зразок програмного забезпечення для апробації отриманих результатів.

Для формування релевантного інформаційного ресурсу на базі статей Wikipedia відповідно до цільового поняття користувача по-перше слід здійснити відбір понять, що відповідають тематиці. Слід відзначити, що кожна стаття відкритої енциклопедії містить досить багато посилань на інші статті загального типу, що напряду не стосуються предметної області поняття. Це можуть бути посилання на мову, рік та ін. поняття, які не повинні включатися в результуючий набір статей. На основі спостереження за структурою статей було розроблено правило, що використовує взаємні цитування, що в сукупності формують певну множину тематично споріднених понять. Наведемо визначення.

- $rel(c_1, c_2)$ – поняття c_1 та c_2 зв'язані за змістом.
- $Link = C \rightarrow C : \{(c_i, c_k)\}$ – множина пар таких понять, що існує посилання з статті поняття c_i в статтю c_k .
- Під нульовим рівнем будемо розуміти цільове поняття, що є навчальним інтересом користувача, та поступає на вхід системи – c_0 .
- $Level_1(c_0) = \{C : (c_0, c) \in Link\}$ – множина понять, що належать першому рівню вкладеності статей.
- $Level_2(c_0) = \{c_k \in Level_1(c_0) \wedge (c_k, c) \in Link\}$ – другий рівень вкладеності статей. Аналогічним чином здійснюється відбір понять третього рівня:
- $Level_3(c_0) = \{c_k \in Level_2(c_0) \wedge (c_k, c) \in Link\}$ – третій рівень вкладеності статей.

У даній роботі для отримання результату використовувалось три рівня вкладеності, але у подальшому розвитку системи планується збільшення кількості рівнів. Збільшення рівнів вкладеності буде сприяти більшому набору понять для обробки та дозволить побудувати більш розгалужену та детальну мапу поняття, заданого користувачем.

Правило визначення споріднених понять для другого рівня передбачає перевірку, чи на сторінці поняття другого рівня вкладеності зустрічається посилання на поняття нульового

або першого рівня. Якщо перевірка дала позитивний результат, тоді ці поняття вважаються тематично спорідненими з цільовим.

Формальний запис правила визначення тематичної спорідненості понять №1:

$$c \in \text{Level}_2(c_0) \wedge ((c, c_0) \in \text{Link}) \rightarrow \text{rel}(c_0, c) \langle \text{CF}_{20} \rangle,$$

де c_0 – вхідне поняття, CF_{20} – фактор впевненості, що вказує на ступінь достовірності висновку відповідно до моделі Б'юкенона[2].

Правила визначення тематичної спорідненості понять №2:

$$c \in \text{Level}_2(c_0) \wedge \exists c_k: (c_k \in \text{Level}_1(c_0) \wedge (c, c_k) \in \text{Link}) \rightarrow \text{rel}(c_0, c) \langle \text{CF}_{21} \rangle,$$

де c_0 – вхідне поняття.

Для правила дидактичного порядку на основі назви понять використовується логіка, запропонована в [1]. Синтаксична наявність у назві поняття «1» назви поняття «2» є аргументом на користь того, що поняття «1» потрібно вивчати раніше ніж поняття «2», таким чином стаття «2» слідує за статтю «1» у ланцюжку вивчення.

Важливим фактором для послідовності вивчення понять є позиція входження посилань. Чим менша позиція, на якій було знайдено посилання на інше поняття, тим більша впевненість у тому, що знайдене поняття слід вивчати раніше.

Якщо поняття «1» фігурує у наборі посилань, які були знайдені на сторінці поняття «2», то поняття «1» є дидактичною передумовою поняття «2» з деяким ступенем достовірності, що обернено пропорційно залежить від позиції входження посилання «1» на сторінці статті поняття «2». На рисунку 4 схематично зображено приклад посилань, які знайдені на різних позиціях входження.

ВИСНОВКИ. В роботі проведено аналіз структури статей Wikipedia та зв'язків між ними на базі гіперпосилань. Розроблено логічні правила визначення тематичної спорідненості понять, що ґрунтуються на використанні трьох рівнів вкладеності статей за посиланнями. Дані правила дозволили автоматично відбирати тематично пов'язані статті за цільовим поняттям.

Запропоновано правила визначення дидактичного порядку між статтями Wikipedia. Правило за іменем, представлене в роботі [1] було успішно застосовано до статей відкритої енциклопедії. На основі аналізу особливостей структури статей Wikipedia було розроблено правило визначення дидактичного порядку статей за позицією входження.

Сукупність запропонованих правил визначення тематичної спорідненості та дидактичного порядку дозволили сформулювати апарат нечіткого виведення для автоматичного формування інформаційно-навчального ресурсу на базі Wikipedia для заданого цільового поняття.

Запропонований формальний апарат реалізовано у програмній системі, що дозволило провести досліду апробацію розроблених логічних правил. Отримані результати свідчать про перспективність запропонованого апарату.

Наступні дослідження будуть зосереджені на удосконаленні програмного забезпечення, проведені ширших дослідних випробувань та виявленні додаткових закономірностей в інформаційних ресурсах Wikipedia, що дозволить удосконалити апарат логічного виведення для вирішеної задачі.

ЛІТЕРАТУРА

1. Титенко С. В. Онтологически-ориентированная система управления контентом информационно-учебных Web-порталов / С. В. Титенко // Educational Technology & Society — 15 (3). 2012. — pp. 522-533. ISSN 1436-4522
 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.
-