УДК 004.896

СИСТЕМА АВТОМАТИЗОВАНОЇ ПОБУДОВИ НАВЧАЛЬНИХ РЕСУРСІВ НА ОСНОВІ СТАТЕЙ WIKIPEDIA

І. А. Супряга, С. В. Титенко
Національний технічний університет України «Київський політехнічний інститут»
просп. Перемоги, 37, м. Київ, 03056. E-mail: supryaga.igor@gmail.com

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. Сукупність запропонованих правил дозволили сформувати 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 дослідну апробацію розроблених нечітких правил. Отримані результати свідчать про перспективність запропонованого апарату.

Ключові слова: Wikipedia, логічне виведення, нечітка логіка, апарат Б'юкенона, дидактичний порядок понять, інформаційно-навчальні ресурси, автоматизовані навчальні системи.

СИСТЕМА АВТОМАТИЗИРОВАННОГО ПОСТРОЕНИЯ ОБУЧАЮЩИХ РЕСУРСОВ НА ОСНОВЕ СТАТЕЙ WIKIPEDIA

И. А. Супряга, С. В. Титенко
Национальный технический университет Украины «Киевский политехнический институт»
просп. Победы, 37, г. Киев, 03056, Украина. E-mail: supryaga.igor@gmail.com

Проведен анализ структуры статей Wikipedia и связей между ними. Разработаны логические правила определения тематической связи понятий, основанные на использовании трех уровней вложенности. Предложены правила определения дидактического порядка между статьями Wikipedia. На основе анализа особенностей структуры статей Wikipedia было разработано правило определения дидактического порядка статей по позиции вхождения. Успешно применено правило определения дидактического порядка по имени к статьям Wikipedia. Совокупность предложенных правил позволили сформировать аппарат нечеткого вывода для автоматического формирования информационно-учебного ресурса на базе Wikipedia для заданного целевого понятия. Предложенный формальный аппарат реализован в программной системе, что позволило провести исследовательскую апробацию разработанных нечетких правил. Полученные результаты свидетельствуют о перспективности предложенного аппарата.

Ключевые слова: Wikipedia, логический вывод, нечеткая логика, аппарат Бьюкенона, дидактический порядок понятий, информационно-обучающие ресурсы, автоматизированные обучающие системы.

АКТУАЛЬНІСТЬ РОБОТИ. На даний момент у всесвітній павутині розміщено величезні об'єми інформації навчального спрямування. Доступність навчальної інформації стає перспективним підґрунтям для самонавчання. З іншого боку інформаційна перенасиченість породжує необхідність докладання значних зусиль для впорядкування даних та відбору статей, що відповідають навчальним інтересам користувача. Таким чином, актуальною задачею для навчальних ресурсів та систем є подання інформаційних матеріалів у зручній послідовності з урахуванням дидактичних взаємозв'язків, що скоротило б час на пошук релевантної навчальної інформації.

Великий об'єм інформації навчального спрямування міститься на порталі Wikipedia, структуру і вміст якого користувачі можуть самостійно змінювати за допомогою інструментів, які надаються системою. Форматування та вставка різноманітних об'єктів у текст відбувається за допомогою вікі-розмітки. На 24 грудня 2013 року цей навчальний ресурс містив більше 30 млн статей, що написані добровольцями з усього світу. Як інтернет-довідник Wikipedia є найбільшою і найпопулярнішою серед подібних сайтів. За обсягом відомостей і тематикою вона вважається найповнішою енциклопедією, яка коли-небудь створювалася за всю історію людства. У зв'язку з цим робота з Wikipedia як базою знань та сховищем навчального контенту є актуальною і практично доцільною задачею. Кожна стаття у своєму тексті містить певні поняття, що вже описані в енциклопедії. Такі поняття подаються у вигляді посилань на статті Wikipedia. Вочевидь, між такими статтями існує смисловий та дидактичний зв'язок. Архітектура Wikipedia передбачає доступ до вихідної вікі-розмітки, що робить можливим автоматичний аналіз текстової інформації кожної із статей.

Метою даної роботи є дослідження та розробка методів автоматичної побудови інформаційно-навчальних ресурсів на базі Wikipedia відповідно до навчальної мети користувача.

Постановка задачі. Під навчальною метою розуміється деяке вхідне поняття – стаття енциклопедії. Цільовий навчальний ресурс – сукупність упорядкованих статей, пов'язаних із цільовою статтею. Такі статті є дидактичними попередниками цільового поняття. 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.

На сьогодні немає системи, яка працює з Wikipedia та має необхідні в контексті даної роботи функціональні можливості. Натомість існує програмний засіб, що функціонує на інформаційно-навчальному порталі znannya.org, який реалізує подібні функції. Впорядкування контенту на сайті здійснюється на основі відношень між статтями. Статті мають зв'язки між собою, що сприяє легкому пошуку інформації користувачем та спрощенню керування контентом для власників сайту. Користувач, що потрапляє на ресурс znannya.org, цікавлячись однією темою знаходить матеріал, що пов'язаний з його запитом, та йому не потрібно повторно здійснювати пошук інформації в пошукових системах чи на інших ресурсах, що скорочує час самонавчання. Впорядкування дидактичних матеріалів порталу відбувається на основі логічних правил [1, 2] із застосуванням стенфордської моделі нечіткого виведення Б'юкенона [3].

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. Методи визначення дидактичного порядку понять в [2] ґрунтуються на додатковій інформації в БЗ про навчальні поняття, яка зберігається в невеликих фрагментах тексту природною мовою відповідно до понятійно-тезисної моделі (ПТМ). Ці фрагменти називають тезами. Теза, як правило, подається у вигляді одного речення, в якому іде мова про поняття, до якого дана теза стосується. Синтаксичний аналіз тез дозволяє виявити згадування інших понять в описі певного поняття предметної області. Натомість, на основі згадувань не можна точно стверджувати про дидактичний порядок. Тому в [2] застосовується апарат нечіткого виведення Б'юкенона, адаптований для даної задачі.

Предикат, що служить для опису твердження про те, що деяке поняття c_k дидактично передує деякому поняття c_l:

$$concept\_before(c_k, c_l) \ .$$

На базі багатьох фактів про взаємне згадування понять в тезах, а також в назвах самих понять, відпрацьовує система нечітких логічних правил, що дозволяє обрати дидактичний зв'язок, що є найбільш достовірним. Для кожного правила експертом задається фактор впевненості *CF*, що вказує на ступінь достовірності. Після спрацювання усіх правил про гіпотезу $concept\_before(c_k, c_l)$ формується множина факторів впевненості $CFs(c_k, c_l)$, які об'єднуються за формулою [2, 3]:

$$CF = CF + CF_i - CF \cdot CF_i \ ,$$

де $CF_i \in CFs(c_k, c_l)$, $i=2..n$. На першій ітерації $CF=0$.

Базовими правилами для визначення дидактичного слідування понять в ПТМ є наступні [2]:

– Правило ПТМ № 1. Якщо поняття «1» фігурує в назві поняття «2», то поняття «1» дидактично передує поняттю «2» з високим ступенем достовірності.

– Правило ПТМ № 2. Якщо поняття «1» фігурує в тезі поняття «2», то поняття «1» дидактично передує поняттю «2» з деякою достовірністю.

– Правило ПТМ № 3.Також для деяких випадків діятиме зворотне правило: якщо поняття «1» фігурує в тезі поняття «2», то поняття «2» дидактично передує поняттю «1» з деякою достовірністю.

Запропонований апарат визначення дидактичного порядку використовується на порталі znannya.org для побудови дидактичної онтології, візуалізації дидактичних карт понять, а також для подальших розрахунків для визначення дидактичного порядку між цілими ділянками навчального контенту та навчальними курсами [1, 2].

Правила [2] не можна напряму застосувати до Wikipedia, так як вони ґрунтуються на використанні структури ПТМ, яка відсутня в енциклопедії. З іншого боку, правило ПТМ № 1 спирається лише на назву поняття (без тез) і тому може бути розглянуте для безпосереднього використання для впорядкування статей Wikipedia, кожна з яких подає інформацію перш за все про конкретне поняття. Правила ПТМ № 2 та № 3 не можуть бути застосовані напряму та потребують додаткового вивчення та можливої адаптації для використання в межах дидактичного впорядкування статей Wikipedia.

Таким чином система, яка відповідає за впорядкування контенту сайту znannya.org, може обробляти лише інформацію, що зберігається в базі знань даного навчального ресурсу та не пристосована для обробки статей відкритої енциклопедії Wikipedia. Відмінність інформаційної структури znannya.org [1] та Wikipedia не дозволяє безпосередньо застосувати апарат логічного виведення znannya.org для визначення дидактичного порядку статей відкритої енциклопедії.

Визначення споріднених понять. 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 напряму 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. Поняття c_1 та c_2 пов'язані за змістом:

$$\mathrm{rel}(c_1, c_2).$$

Множина пар таких понять, що існує посилання з статті поняття c_i в статтю c_k:

$$\mathrm{Link} = C \to C : \{(c_i, c_k)\}.$$

Визначення спорідненості спирається на поняття рівнів вкладеності (рис. 1). На рис. 1 вершинам відповідають статті енциклопедії, а ребрам – гіперпосилання.

Під нульовим рівнем будемо розуміти цільове поняття, що є навчальним інтересом користувача, та поступає на вхід системи – c_0.

Множина понять, що належать першому рівню вкладеності статей:

$$\mathrm{Level}_1(c_0) = \{C: (c_0, c) \in \mathrm{Link}\}.$$

Другий рівень вкладеності статей:

$$\mathrm{Level}_2(c_0) = \{c: c_k \in \mathrm{Level}_1(c_0) \wedge (c_k, c) \in \mathrm{Link}\}.$$

Аналогічним чином здійснюється відбір понять третього рівня:

$$\mathrm{Level}_3(c_0) = \{c: c_k \in \mathrm{Level2}(c_0) \wedge (c_k, c) \in \mathrm{Link}\}.$$

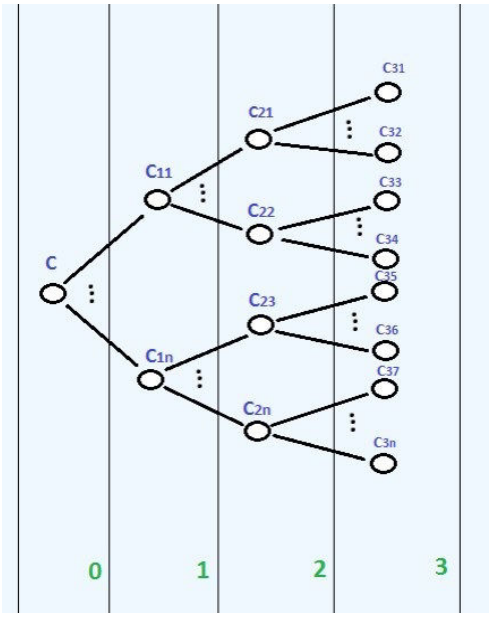

Рисунок 1 – Схема поділу понять на рівні вкладеності

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:

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

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:

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

де c_0 – вхідне поняття.

На рис. 2 схематично зображено приклад набору понять, коли поняття другого рівня вкладеності посилаються на поняття першого або нульового рівня вкладеності.

Аналогічно, якщо на сторінці поняття третього рівня вкладеності зустрічається посилання на поняття нульового, першого або другого рівня, то ці поняття є тематично спорідненими з деяким ступенем достовірності. На рис. 3 схематично зображено приклад набору понять, коли поняття третього рівня вкладеності посилаються на поняття другого, першого або нульового рівня.

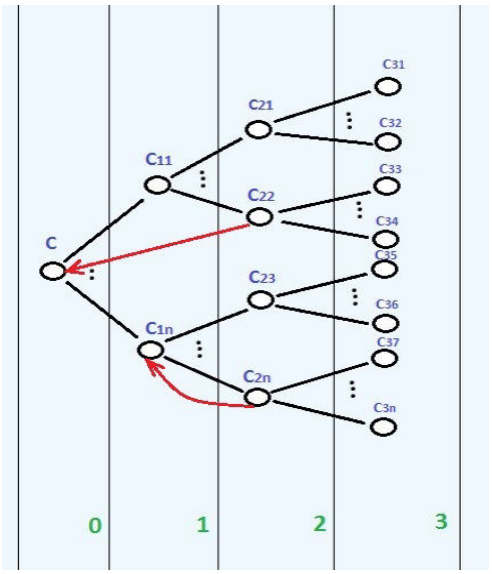

Рисунок 2 – Приклад набору понять, коли поняття другого рівня вкладеності посилаються на поняття першого або нульового рівня вкладеності

Відповідно, правило № 3 визначення спорідненості понять формалізується наступним чином:

$$c \in \mathrm{Level}_3(c_0) \wedge \big((c, c_0) \in \mathrm{Link}\big) \longrightarrow \mathrm{rel}(c_0, c)\langle \mathrm{CF}_{30}\rangle$$

Правило № 4 визначення спорідненості понять для третього рівня:

$$C \in \mathrm{Level}_3(c_0) \wedge \exists c_k: (c_k \in \mathrm{Level}_1(c_0) \wedge (C, c_k) \in \mathrm{Link}) \rightarrow \mathrm{rel}(c_0, c)\langle \mathrm{CF}_{31}\rangle.$$

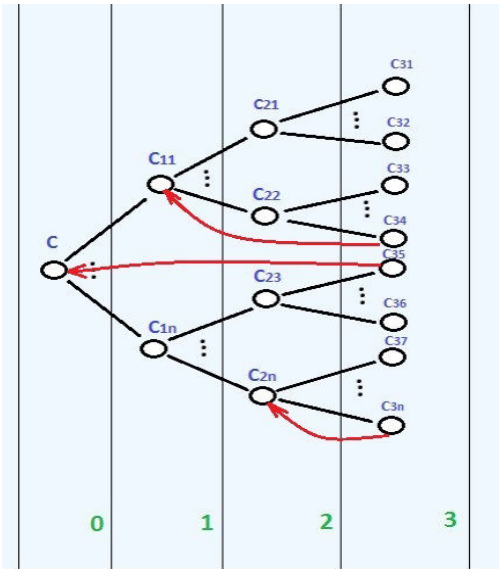

Рисунок 3 – Приклад набору понять, коли поняття третього рівня вкладеності посилаються на поняття другого, першого або нульового рівня вкладеності

Правило № 5 визначення спорідненості понять для третього рівня:

$$C \in \text{Level}_3(c_0) \wedge \exists c_k: (c_k \in \text{Level}_2(c_0) \wedge (c, c_k) \in \text{Link}) \longrightarrow \text{rel}(c_0, c)\langle CF_{32}\rangle.$$

Запропоновані логічні правила дозволяють відібрати такі статті Wikipedia, які за змістом відповідають тематиці вхідного цільового поняття, вказаного користувачем.

Визначення дидактичного порядку понять за іменем. Після відбору споріднених понять перед системою стоїть задача впорядкувати статті таким чином, щоб вони відповідали дидактичному порядку вивчення понять. Адже для засвоєння інформації про деяке поняття в Wikipedia користувач повинен, використовуючи навігаційні посилання в тексті, здійснити перехід та ознайомлення з іншими поняттями, на яких ґрунтується дане. Нашим завданням є передбачити, які з понять є дидактичними попередниками цільового поняття, та автоматично побудувати навігаційну схему вивчення, тим самим спростивши роботу користувачу.

Для правила дидактичного порядку на основі назви понять використовується логіка, запропонована в [2]. 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 статею «1» у ланцюжку вивчення.

Наприклад, на сайті Wikipedia є стаття «Абстрактний клас», та стаття з назвою «Клас». У назві першого поняття зустрічається назва другого, тобто початок дидактичної послідовності – це стаття «Клас», а кінець – стаття «Абстрактний клас».

Правило можна формально записати у наступному вигляді:

$$c_k \in \text{CinC}(c_i) \rightarrow \text{concept\_before}(c_k, c_i)\langle CF_{\text{cinc}}\rangle,$$

де $\text{concept\_before}(c_k, c_i)\langle CF_{\text{cinc}}\rangle$ означає, що поняття c_k є дидактичним попередником поняття c_i з ступенем достовірності $\langle CF_{\text{cinc}}\rangle$,

$c_k \in \text{CinC}(c_i)$ – у назві поняття c_i знайдено назву поняття c_k.

Визначення дидактичного порядку понять за позицією. 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 рис. 4 схематично зображено приклад посилань, які знайдені на різних позиціях входження.

При виникненні ситуації, коли два поняття посилаються одне на одне з однаковим ступенем достовірності, здійснюється усунення протиріч на основі кількості входжень. Для цього розбирається текст першого поняття та знаходиться кількість входжень назви статті другого поняття. Потім, аналогічно, розбирається текст другого поняття.

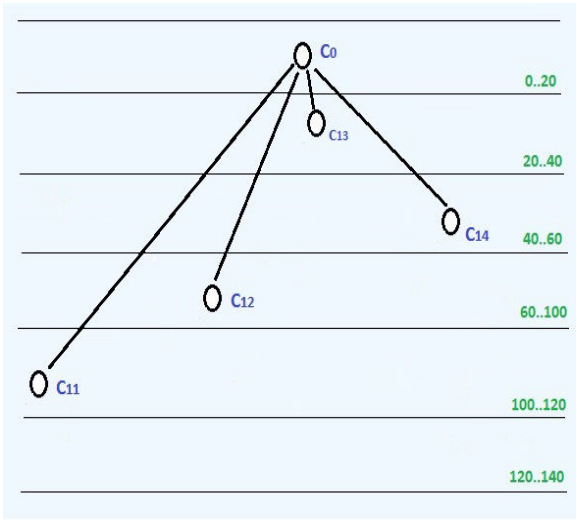

Рисунок 4 – Позиція входження посилання на інше поняття в статті Wikipedia

На рис. 5 схематично зображено приклад протиріччя.

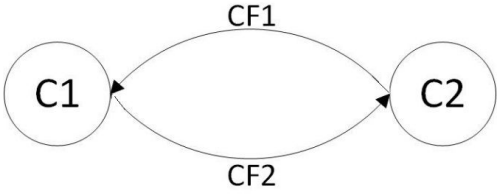

Рисунок 5 – Приклад протиріччя

Правила усунення протиріч на основі кількості входжень формалізовано можна записати у вигляді:

$$\text{concept\_before}(c_k, c_i)\langle CF_{ki}\rangle \wedge$$
$$\text{concept\_before}(c_i, c_k)\langle CF_{ik}\rangle \wedge (CF_{ik} = CF_{ki}) \wedge$$
$$\text{Num}(c_i, c_k) > N \quad c_k, c_i) \rightarrow$$
$$\text{concept\_before}(c_k, c_i)\langle CF_{\text{Num}}\rangle,$$

де $\text{Num}(c_i, c_k)$ — кількість входження поняття c_k у тексті на сторінці поняття c_i.

Програмна реалізація. Запропонований формальний апарат програмно реалізовано за допомогою засобів розробки PHP, MySQL. Програмна система на вході отримує URL статті Wikipedia, наступним етапом здійснює синтаксичний аналіз тексту статті та будує навігаційний граф за гіперпосиланнями до третього рівня вкладеності, що формально представляє множину посилань Link. Даний результат зберігається у відповідних структурах бази даних. На області множини посилань здійснюється логічне виведення по правилам № 1–5, і отримується множина тематично споріднених понять. Для множини споріднених понять виконуються логічні правила визначення дидактичного порядку, що дають в результаті граф тематично споріднених понять з відношеннями дидактичного порядку. Отриманий граф можна представити як лінійну послідовність завдяки застосуванню алгоритмів топологічного сортування графу.

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 [2] 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 сформувати 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 дослідну 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 вирішуваної задачі.

ЛІТЕРАТУРА

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

2. Титенко С.В. Побудова дидактичної онтології на основі аналізу елементів понятійно-тезисної моделі // Наукові вісті НТУУ "КПІ". – 2010. – № 1 (69). – С. 82–87.

3. 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.

COMPUTER-AIDED CONSTRUCTION OF LEARNING RESOURCES BASED ON THE WIKIPEDIA ARTICLES

I. Supryaga, S. Tytenko

National Technical University of Ukraine "Kyiv Polytechnic Institute"

prosp. Peremogy, 37, Kyiv, 03056, Ukraine. E-mail: supryaga.igor@gmail.com

The aim of this work is to investigate and develop methods of automatic construction of information and educational resources on the basis of Wikipedia according to learning goal of the user. Were analyzed the structure of Wikipedia articles and the links between them. Were developed logical rules to determine the domain connection of concepts based on the use of three levels of nesting. Were proposed the rules for determining didactic order between Wikipedia articles. On the basis of structure analyzing of Wikipedia articles has been developed rule for determining the order of didactic articles according to occurrence position. The set of proposed rules allowed to form fuzzy inference apparatus for automatic generation of information and educational resources on the basis of Wikipedia for a given target concept. Proposed formal apparatus is implemented in a software system that enabled the research of testing the fuzzy rules. The proposed formal apparatus implemented with the help of software development tools PHP, MySQL. Software system at the input receives URL of Wikipedia article, the next step performs articles text parsing and builds a navigation graph for hyperlinks to third level of nesting, which formally represents the set of links. This result is stored in the appropriate database structure. The set of inference rules for domain connection is performed on the area of links, thus obtained a set of thematically related concepts. For the set of related concepts logical rules of didactic manner executed, giving as a result the graph of thematically related concepts with didactic connections. The resulting graph can be represented as a linear sequence through the use of algorithms for topological sorting of the graph. The obtained results demonstrate the promise of the proposed system.

Key words: Wikipedia, inference, fuzzy logic, Buchanan, didactic order of learning concepts, information and training resources, e-learning systems, ITS, adaptive educational systems.

REFERENCES

1. Tytenko, S.V. (2012) «Ontologically-oriented content management system for information and educational Web-portals», *Educational Technology & Society*, no. 15 (3), pp. 522–533. ISSN 1436-4522.

2. Tytenko, S.V. (2010) "Construction of Didactic Ontology Based on the Analysis of Concept-Thesis Model Elements", *Naukovi Visti NTUU "KPI"*, no. 1 (69), pp. 82–87.

3. Buchanan, B.G., Shortliffe, E.H. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, MA, Addison-Wesley.

Стаття надійшла 22.10.14.