Формування електронної енциклопедії за допомогою екстракції знань з відкритих текстів

Павло Жежнич[1], Марія Гірняк[2]

1. Кафедра інформаційних систем та мереж, Національний університет "Львівська політехніка", УКРАЇНА, м.Львів, вул.С.Бандери, 12,
E-mail: pzhe@ridne.net

2. Кафедра соціальних комунікацій та інформаційної діяльності, Національний університет "Львівська політехніка", УКРАЇНА, м.Львів, вул.С.Бандери, 12, E-mail: maria.girniak@gmail.com

This article dwells on the knowledge extraction from plaintexts in the context of electronic encyclopedia design.It determines the basic steps in knowledge extraction and defines the peculiarities of electronic encyclopedias.

Ключові слова – електронна енциклопедія, екстракція знань, шаблони, макроструктура тексту, перехресні відсилання, відкритий текст.

I. Вступ

У створенні електронної енциклопедії важливим завданням є її інформаційне наповнення, що становить собою непростий процес із використанням різноманітних методів. Так, екстракція знань уможливлює порівняння, аналіз та синтез різнорідної інформації, що міститься у відкритих текстах, доступних широкому загалу та призначені для зберігання, передачі та перетворення.

II. Основна частина

Екстракція знань – формулювання знань із структурованих (реляційні бази даних, XML) та неструктурованих (тексти, документи, зображення) джерел. Отримані знання подаються у формі, придатній для машинного читання та інтерпретації і відображають логічне завершення тексту [4].

Серед екстракції знань виділяють комунікативні та текстологічні методи. Розглядаючи концепцію наповнення електронної енциклопедії, значимості набувають текстологічні методи екстракції знань, які у свою, чергу містять: аналіз підручників, технічної літератури та аналіз спеціалізованої літератури.

Алгоритм екстракції знань із відкритих текстів для формування електронної енциклопедії можна представити наступною послідовністю кроків:

1. Сформувати «базовий» список літератури з огляду на тематичні рубрикатори та тематичні цикли у межах певної галузі знань із чітким визначенням предметної області.

2. Обрати науково-вивірені відкриті тексти для екстракції знань з огляду на те, що енциклопедичні статті мають відповідати наступним вимогам:

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

- науковість (подання матеріалу із достовірних джерел, відсутність особистих прогнозів та гіпотез);
- доступність (відсутність псевдонаукового; термінологічно перенасиченого стилю);
- актуальність (за рахунок постійного оновлення та доповнення);
- повнота (вичерпність інформації; сприяє максимальній інформативності);
- об’єктивність викладу матеріалу;
- подання інформації без індивідуально-емоційного забарвлення;
- наявність науково-довідкового апарату (вказівники, бібліографічні та етимологічні довідки);
- фактологічна точність [1, 3].

3. Побіжне знайомство з відкритим текстом. З’ясувати значення невідомих слів (спеціалізованої термінології), використовуючи словники та енциклопедії.

4. Сформувати макроструктуру відкритого тексту, створити шаблони для представлення знань.

5. Уважно прочитати відкритий текст та визначити ключові слова і вирази, що попередньо забезпечуватиме систему перехресних відсилань.

6. Визначити зв'язки між ключовими словами, розробити макроструктуру відкритого тексту.

7. Сформувати нове подання знань на підставі макроструктури відкритого тексту [2].

Висновок

Створення електронної енциклопедії головним чином базується на екстракції знань, здебільшого із відкритих текстів. Подальшого вивчення потребують способи автоматизації екстракції знань, розроблення шаблонів для подання інформації та питання критеріїв оцінювання якості розроблюваної електронної енциклопедії.

Література

1. Карпіловська Є. Енциклопедія «Українська мова»: структура та принципи укладання / Карпіловська Є., Зяблюк М. // Енциклопедичний вісник України. – К., 2009. – Число1. – С. 47- 53.
2. Методы извлечения знаний [Електронний ресурс]. – Режим доступу: http://asp.mmc.nsu.ru/default.aspx?db=book_zagorulko&int=VIEW&el=1777&templ=I206
3. Черниш Н. І. Українська енциклопедична справа: історія розвитку, теоретичні засади підготовки видань. / Н. І. Черниш. – Львів: Фенікс, 1998. – 92 с.
4. Knowledge extraction [Electronic resource]. – Mode of access: http://en.wikipedia.org/wiki/Knowledge_extraction

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua