UDK 004.9

P. Zhezhnych, M. Hirnyak
Lviv Polytechnic National University,
Information Systems and Networks Department

THE ARCHITECTURE OF THE ELECTRONIC ENCYCLOPEDIA FORMATION SYSTEM ON THE BASIS OF OPEN TEXTS

This paper dwells on the problems of the architecture of the electronic encyclopedia formation system, encyclopedic entry (article) information content and the process of knowledge extraction from open texts.

Key words: electronic encyclopedia architecture, encyclopedic entry (article), formation system, knowledge extraction, information retrieval, open text.

Розглянуто проблеми архітектури системи формування електронної енциклопедії, інформаційного наповнення енциклопедичних статей та процесу екстракції знань з відкритих текстів.

Ключові слова: архітектура електронної енциклопедії, енциклопедична стаття, система формування, екстракція знань, інформаційний пошук, відкритий текст.

Introduction

Nowadays there is a lot of information concerning that or other topic in the Internet space. Search engine (Google, Yandex) receives an inquiry and generates data containing a lot of unnecessary and insufficient information. This information is given in the form of text documents, diagrams, graphs, audio – and videomaterials in web-portals, forums, encyclopedias, etc. In most cases to get a sound, terse and reliable information people search in the electronic encyclopedias to quickly get the answer to their inquiry. Moreover, during the last decade it is observed a rapid grow of the encyclopedias in the World Wide Web. For the developers, the issue to be studied is the electronic encyclopedia formation system with the limited time for the process of its information content; it is not an easy task that demands to apply different methods. It is possible providing the knowledge extraction from open texts with the analysis, synthesis and comparison of a wide spectrum of information that is stored in the open texts.

Information extraction for the electronic encyclopedia information content

In the process of electronic encyclopedia information content there are 2 ways to provide the requisite information:

- information retrieval (input – search for the text in the search engine; output – a set of full texts in the interested area containing likely matches);
- information extraction (input – the analysis of the texts; output – extraction of the information snippest in the form of the fixed-format, unambiguous data).

The difference between information retrieval and information extraction is very essential as far as after the information retrieval the developer of the electronic encyclopedia must extract the relevant requisite information him/herself. Information extraction provides, to some extent, the semi-automated process of information content. However, there are some advantages and disadvantages with information extraction in comparison to information retrieval. As for the advantages, it should be mentioned the following: a sound and appropriate reduction of the text information, selecting only the necessary information and, in its turn, reducing the amount of time to be needed read the full texts; suitable processing concerning the multilingual texts. Nevertheless, information extraction has some disadvantages: systems are more difficult and knowledge-intensive to build; more computationally intensive than information retrieval [1].

The process of the information extraction for the encyclopedic entry content could be represented in the following way (fig. 1):

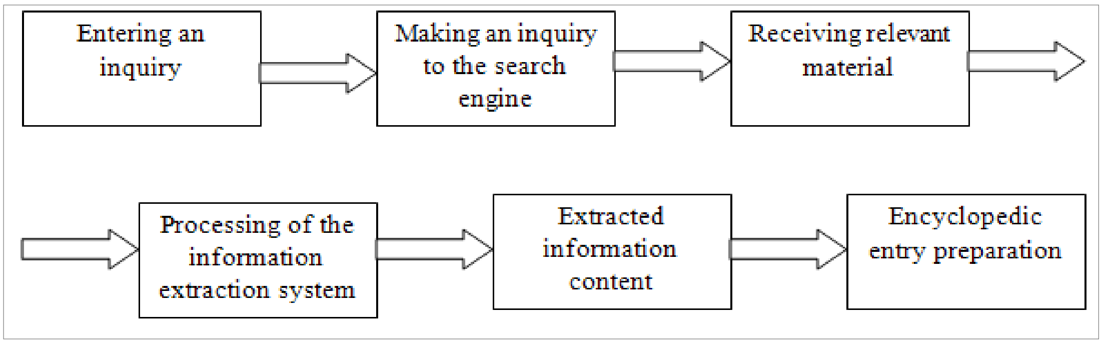
Entering an inquiry
Making an inquiry to the search engine
Receiving relevant material
Processing of the information extraction system
Extracted information content
Encyclopedic entry preparation

Fig. 1. The architecture of the information extraction

In science there is a perception that the information extraction should act in the following way:

- find and classify names, places (named entity recognition);
- define identity relations between entities (coreference resolution);
- add descriptive information to named entity results (template element construction);
- find relations between template element entities (template relation construction);
- fit template element and template relation results into specified event scenarios (scenario template production) [1].

The article [3] considers that there are such methods of information extraction:

- hand-coded or learning-based;
- rule-based or statistical.

In the article [2] the architecture of the information extraction system is proposed (fig. 2):

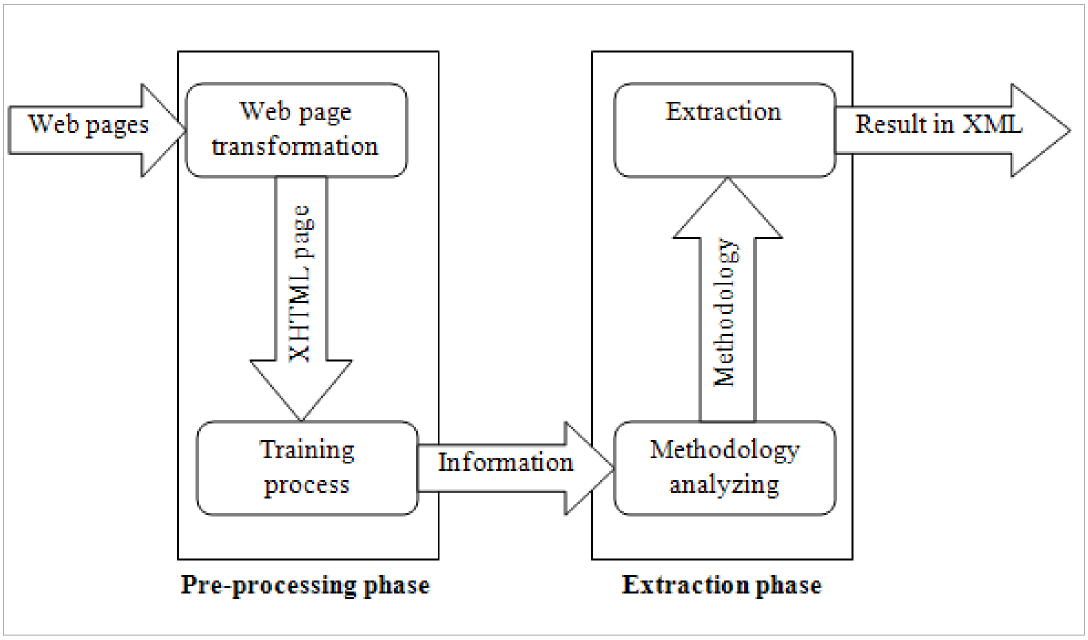

Fig. 2. The information extraction system architecture [2, p.386]

Electronic encyclopedia system formation

In general, having a data flow on the Internet as the input, the developer must design an electronic encyclopedia containing a lot of encyclopedic entries (articles) with the system of cross-references that make up a thematic cycle and hyperlinks to the primary Internet-sources as the output. In addition, an electronic encyclopedia must include the main menu with the detailed system of classifiers (categories) and the internal retrieval system. Therefore, it is obvious that it is impossible to make the formation process of electronic encyclopedia completely automated. Data development, in particular: main menu, classifiers, retrieval system is appointed unto man, while computer helps to select the necessary material for the encyclopedic entries highlighting the essential components of this material owing to the information extraction.

Thus, the architecture of the electronic encyclopedia system formation should be as follows:

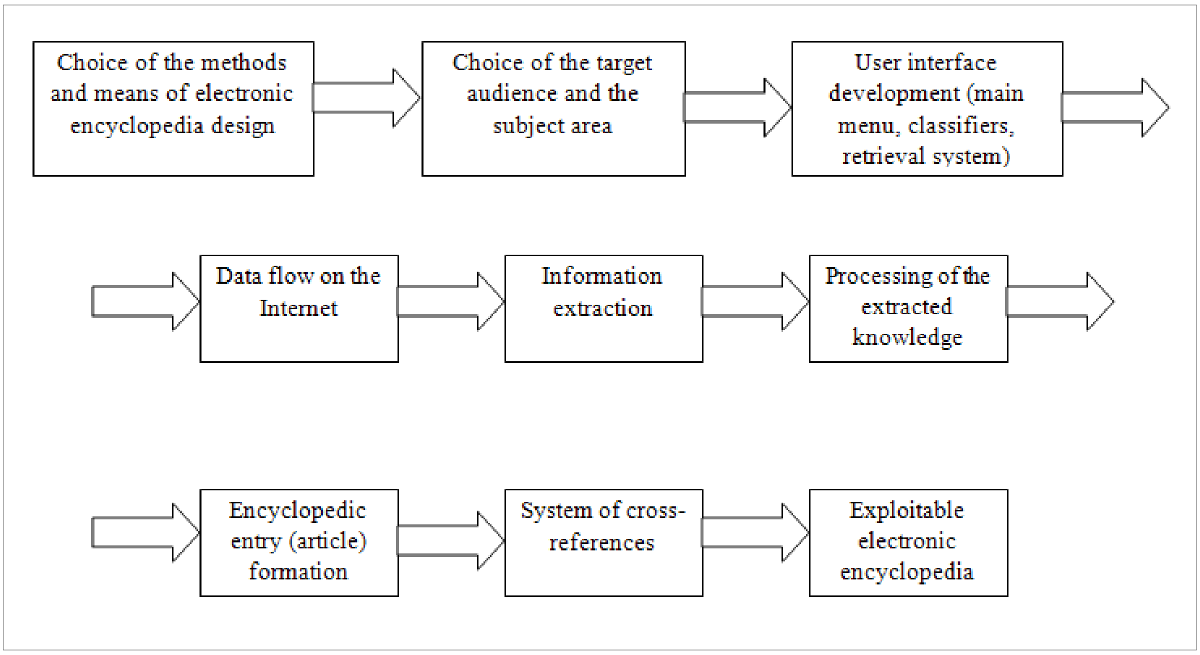
Choice of the methods and means of electronic encyclopedia design
Choice of the target audience and the subject area
User interface development (main menu, classifiers, retrieval system)
Data flow on the Internet
Information extraction
Processing of the extracted knowledge
Encyclopedic entry (article) formation
System of cross-references
Exploitable electronic encyclopedia

Fig. 3. The architecture of the electronic encyclopedia system formation

The electronic encyclopedia improvement is possible with the development of the didactic apparatus. The didactic apparatus consists of the information presentation apparatus, orientation apparatus, apparatus of the information mastering, processing apparatus and the bibliographic apparatus.

The information presentation apparatus contributes to a better visual presentation and more informative material and includes: hypertext, animation, audio – and videomaterials, elements of virtual reality, demonstrational and manipulative dynamic models of the objects and processes. It is the essential attributes for the users with different pathologies (hearing, vision, dyslexia).

As for the orientation apparatus, it is a combination of different indicators, namely: alphabetical, systematic and bibliographic; retrieval system by a keyword, specific term.

The apparatus of the information mastering is characterized by a systematic presentation of the material, taking into account the better visual perception of the information with the help of various schemes, graphs, tables, charts and classifications.

The processing apparatus includes the processes of selection, sorting, categorizing of the information, its statistical processing and editing. This module, in contrast to the previous three, is peculiar only to the electronic encyclopedias.

The last attribute of the electronic encyclopedia didactic apparatus is the bibliographic apparatus. It provides the direction to the reader to the original source of the quoted text and recommends the references for the detailed and sound coverage of the issue or problem discussed in the particular encyclopedic entry [4].

In this way, the process of electronic encyclopedia formation is a semi-automated process. There is a close interaction of the developer and computer.

Information content is very essential for the users and the reputation of the electronic encyclopedia. To provide the logical, consistent and interdependent material presentation the developer should elaborate a system of cross-references and hyperlinks to the primary sources and the additional bibliography guide. It could be represented in the following way (fig. 4):

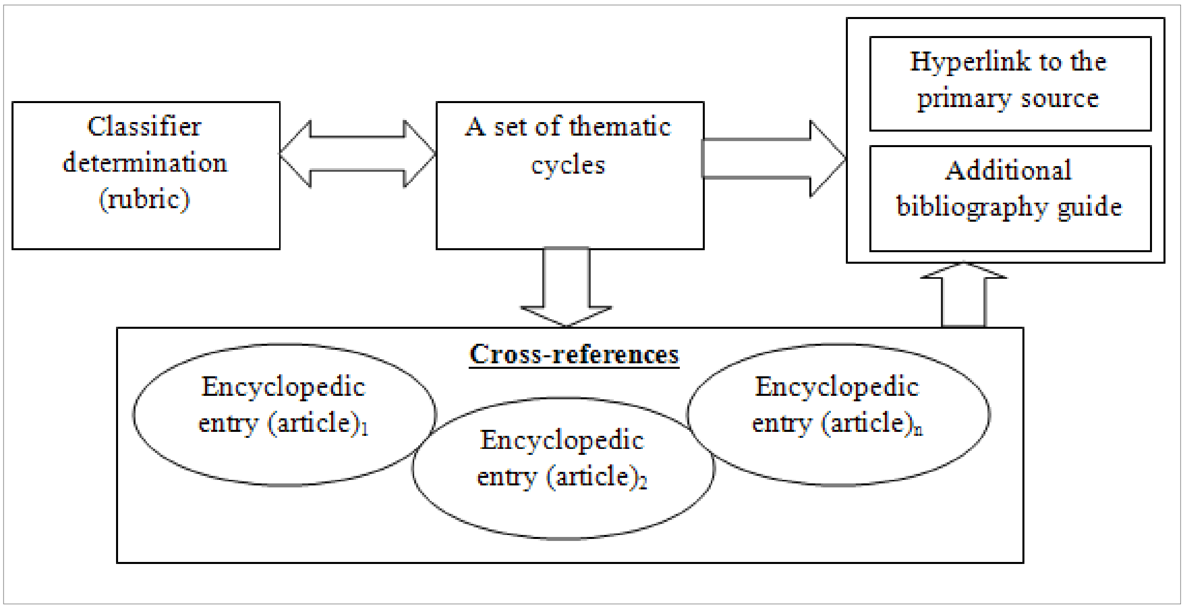
Classifier determination (rubric)
A set of thematic cycles
Hyperlink to the primary source
Additional bibliography guide
Cross-references
Encyclopedic entry (article)₁
Encyclopedic entry (article)₂
Encyclopedic entry (article)ₙ

Fig. 4. A system formation of the encyclopedic information content

Conclusions

Electronic encyclopedia formation requires the systematization of the material in accordance with the developed classifiers, defining the key points and the appropriate cross-references and hyperlinks. The process of information extraction makes the process of electronic encyclopedia formation half-automated. Nevertheless, the architecture of its system formation requires much efforts, time and computer skills. Such issues as the elaboration of the approach to improve the quality of the electronic encyclopedia, elaboration of the electronic encyclopedia information model need further study.

1. Cunningham H. Information Extraction, Automatic / H. Cunninghum // Encyclopedia of Language & Linguistics / Editor-in-chief Keith Brown. – Oxford: Elsevier, 2006. – Second Edition. – Vol. 5. – P. 665–677. 2. Man I Lam A Method for Web Information Extraction / Man I Lam, Zhiguo Gong, and Maybin Muyeba // Proceedings of the 10th Asia-Pacific Web Conference, APWeb: Lecture Notes in Computer Science. – Shenyang, China: Springer, 2008. – P. 383–394. 3. Sarawagi S. Information Extraction / S. Sarawagi // Foundations and Trends in Databases. – Bombay, India: CSE, 2007. – Vol. 1, № 3. – P. 261–377. 4. Жежнич П.І. Особливості формування енциклопедії в сучасних умовах розвитку інформаційних технологій / П.І. Жежнич, М.Г. Гірняк // Вісник Національного університету «Львівська політехніка»: Комп'ютерні науки та інформаційні технології. – Львів: Вид-во Нац. ун-ту «Львівська політехніка», 2012. – №732. – С. 399–405. 5. Кузнецов И. П. Автоматическое формирование электронных энциклопедий и справочных пособий по информации из сети "Интернет" / И. П. Кузнецов, М. М. Шарнин // Системы и средства информатики: ежегодник. – М.: Наука, 2001. – Вып. 14. – С. 210–223.