UDC 004.774

SYSTEM FOR WEB RESOURCES CONTENT STRUCTURING AND RECOGNIZING WITH THE MACHINE LEARNING ELEMENTS

Dyvak M. P. – Dr.Sc., Professor, Dean of the Faculty of Computer Information Technology, Ternopil National Economic University, Ternopil, Ukraine.

Kovbasistyi A. V. – Post-graduate student at the Computer Science Department, Ternopil National Economic University, Ternopil, Ukraine.

Melnyk A. M. – PhD, Associate Professor at the Computer Science Department, Ternopil National Economic University, Ternopil, Ukraine.

Turchyn L. Y. – PhD, Associate Professor at the Business, Trade and Marketing Department, Ternopil National Economic University, Ternopil, Ukraine.

Martsenyuk Y. O. – PhD, Associate Professor at the Computer Science Department, Ternopil National Economic University, Ternopil, Ukraine.

ABSTRACT

Context. A large number of web resources of different organizations requires checking of relevance and correctness of the content, in particular, concerning characteristics of the organization, staff, etc. For this, it is necessary to develop a system of the automated content analysis. This task causes the need to develop a method and software for structuring and recognizing of web resources content. Existing parsing systems do not provide solving of the specified task, since they do not contain elements of machine learning. The object of the research is the process of automated analysis of the web resources content.

Objective. The goal of the work is the creation of the system for web resources content structuring and recognizing.

Method. The system of structuring and recognizing of text content of web resources with elements of machine learning is considered. Models of the system functioning are proposed. The architecture for realizing of software system for structuring and recognizing of text content of web resources is developed. Example of implementation of the model of developed system for structuring, recognizing and revealing of outdated and incorrect information about personnel on the web resource of educational institution is given.

Results. The developed software may be used by support services in order to update and correct the information content.

Conclusions. The system of structuring and recognizing of content of web resources with the machine learning elements has been considered. The proposed system compared with the known ones, ensures automated content structuring, recognizing of outdated, non-relevant or wrong information. Represented example of the structuring and recognizing of outdated and incorrect information on the website of educational institution confirms the effectiveness of the proposed system.

KEYWORDS: Content analysis, parsing, machine learning.

NOMENCLATURE

$\mathrm{Pr}\,O$ is a parent element and the attribute;

Id_Type is an identifier of type which allows to improve the belonging of an object to a group that already has more common characteristics than Dsc characteristic;

O_Phr is a characteristic of the object analyzed, allows different linguistic representation in the form of phrases;

O_Frm is a word and that define specific characteristics presented in word forms;

O_Bs is the words and their bases;

$IdLg$ is an identifier of language implementation;

$IdBs$ is an identifier of words base;

$WBase$ is a base of the word;

$IdFm$ is an identifier of words form;

$WForm$ is a form of words;

$Id\mathrm{O}$ is an identifier of object;

$IdPh$ is an identifier of features;

$Id\,\mathrm{Pr}\,Bs$ is an identifier of the parental basics of the object;

$KDrO$ is a set of key characteristics;

$IdPg$ is an identifier of the analyzed web page;

IdO is an identifier of the analyzed object;

$\mathrm{Id}\_Type\mathrm{E}$ is a type of sample characteristics.

INTRODUCTION

Structuring and recognizing of text information is one of the areas of intelligent information systems. The main component of such systems is machine learning [1].

Machine learning is a process in which a system processes a large number of examples, detects patterns and uses them to predict the characteristics of new data.

Machine learning deprives the programmer of the need to “explain in details” to the computer how to solve the problem. Instead, the computer learns to find a solution on its own [2].

In this paper, the system for structuring and recognizing of web resource content, which includes the elements of machine learning, is proposed. This system can be used for identification of incorrect and outdated information on websites.

The object of study is the process of automated analysis of web resources content.

The subject of study is the **s**ystem for web resources content structuring and recognizing.

The purpose of the work is the development of method and software for web resources content structuring and recognizing using machine learning elements.

1 PROBLEM STATEMENT

Let us consider a system that performs searching, structuring and recognizing of text information on web resources, recording to a database and comparing it.

DOI 10.15588/1607-3274-2018-3-14

The processed object of this system is the input data in the form of text information. The structure of such information is represented in the form of tree-like relations.

Therefore, the objects that are used to analyze the relevance of their content are modeled using following structure for machine learning:

$$O\_\mathrm{Dsc} = < IdO, \mathrm{Pr}\,O, \mathrm{Id}\_Type, \mathrm{Dsc} >, \quad (1)$$

which includes the object identifiers IdO , parent element identifiers $\mathrm{Pr}\,O$ and the attribute Id_Type that allows to highlight the belonging of an object to a certain group that already has more common characteristics than Dsc characteristic. At this, general object differs from particular ones by the absence of parent element, for which, $\mathrm{Pr\,O} = NULL$.

Each characteristic of the analyzed object allows different linguistic representation in the form of phrases O_Phr . Words that define specific characteristics, are represented by phrases O_Frm and by their bases O_Bs . Phrases are used to represent the concepts to users, and bases are used for automatic identification of equivalence of language representations between analyzed data and reference data. The attributes of introduced notions can be grouped into the following structures for machine learning:

$$O\_Bs = < IdLg, IdBs, WBase >, \quad (2)$$

$$O\_Frm = < IdLg, IdFm, IdBs, WForm >, \quad (3)$$

$$O\_Phr = < IdCn, IdLg, IdPh, IdBs, IdFm, Id\,\mathrm{Pr}\,Bs >. \quad (4)$$

Thus, the input data for system for web resources content structuring and recognizing functionality is represented by the set of tuples (1)–(4).

As a result of transformation of given structures, we obtain a tree-like structure in such form:

$$O\_\mathrm{Dsc} = < IdO, \mathrm{Dsc} >, \quad (5)$$

where O_Dsc is resulting set of text structures, the relevance of content for which is established.

The relevance of obtained results of tree-like structures transformation depends on completeness of available content representation in existing databases that are the basis for comparison. Thus, in the content analysis process, the main database must be renewed using machine learning tools. The problem of defining of indicators of content representation completeness in existing databases is beyond the scope of research represented in the paper.

2 REVIEW OF THE LITERATURE

Machine learning can appear in many guises. We now discuss a number of applications, the types of data they deal with, and finally, we formalize the problems in a somewhat more stylized fashion. The latter is key if we want to avoid reinventing the wheel for every new application. Instead, much of the art of machine learning is to reduce a range of fairly disparate problems to a set of fairly narrow prototypes. Much of the science of machine learning is then to solve those problems and provide good guarantees for the solutions [4, 5].

Machine learning algorithms are organized into taxonomy, based on the desired outcome of the algorithm. Common algorithm types include [6]:

– Supervised learning – where the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input-output examples of the function.

– Unsupervised learning – which models a set of inputs: labeled examples are not available.

– Semi-supervised learning - which combines both labeled and unlabeled examples to generate an appropriate function or classifier.

– Reinforcement learning – where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.

– Transduction – similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.

– Learning to learn – where the algorithm learns its own inductive bias based on previous experience.

The performance and computational analysis of machine learning algorithms is a branch of statistics known as computational learning theory. Machine learning is about designing algorithms that allow a computer to learn. Learning is not necessarily involves consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Thus, many machine learning algorithms will barely resemble how human might approach a learning task. However, learning algorithms can give insight into the relative difficulty of learning in different environments [6].

The components of machine learning tools provide the ability of software systems to analyze automatically the text information in order to structure it, determine incorrect and wrong information [5].

Especially important is to develop the machine-adapted procedures for structuring the text information published on the web resources. Automated analysis and structuring of the web resources content makes it possible to solve complex applied problems in economy, ecology, medicine and more [7, 8].

Today, there are tools that can solve these problems to some extent.

In the paper [9, 10], the systems that allow parsing content of the entire website as well as its particular pages, obtaining data from dynamic pages and uploading

DOI 10.15588/1607-3274-2018-3-14

information in the appropriate format for its comparison are considered.

All these systems have the main drawback – they do not deal with the machine learning.

3 MATERIALS AND METHODS

The characteristics of specific object are understood as tree-like system of meaningful notions (concepts) that describe its certain meaningful properties. Some separated typical characteristics for determination of its belonging to a particular group and generalized characteristics for the whole object are included for this.

Let input content of web resource is represented in the form of tree-like structures (1)–(4). Description of the main information structures is represented in Fig. 1.

For comparison of analyzed characteristics values and their corresponding reference data, it is necessary to analyze the information represented on corresponding web pages [11].

The structure of description of information about analyzed objects is unknown a priori and may vary depending on the subject, content and technical realization of these websites. It is expedient that information on websites must be structured, not just divided into blocks or paragraphs [12]. This requirement can significantly simplify the content relevance analyzing process. In this case, structuring means the representation of information in the form of defined structures.

To analyze information about a particular object, the set of key characteristics $KDrO$ is formed. To support the analysis of the web pages content, the following auxiliary structure AS was created [13, 14]:

$$AS =< IdPg, IdO, IdIt, IdBs, IdFm, Id\Pr Bs >, \qquad (6)$$

During analysis of HTML code of next web page of specialized website, its identifier is set:

$$CurPgId := \max\left(\pi_{IdLPg}(BF)\right) + 1, \qquad (7)$$

Then, we distinguish the elements $LSTIt$ of corresponding characteristics that fill corresponding tags. These elements are used further for comparison with the elements of reference characteristics of objects OE_Dsc of specific database. It is described by the following structure:

$$OE\_Dsc =< IdO, \Pr O, Id\_TypeE, DscE >, \qquad (8)$$

Let $Wrd_k(It_{Pg,Lst})$ is some k-th characteristic distinguished from the corresponding object $It_{Pg,Lst}$. In this case, machine learning procedure is described by such structure:

$$BsWrd = \pi_{Dsc}\left(\sigma_{DscE=Dsc(Wrd_k(It_{Pg,Lst}))}(O\_Dsc, OE\_Dsc)\right), \qquad (9)$$

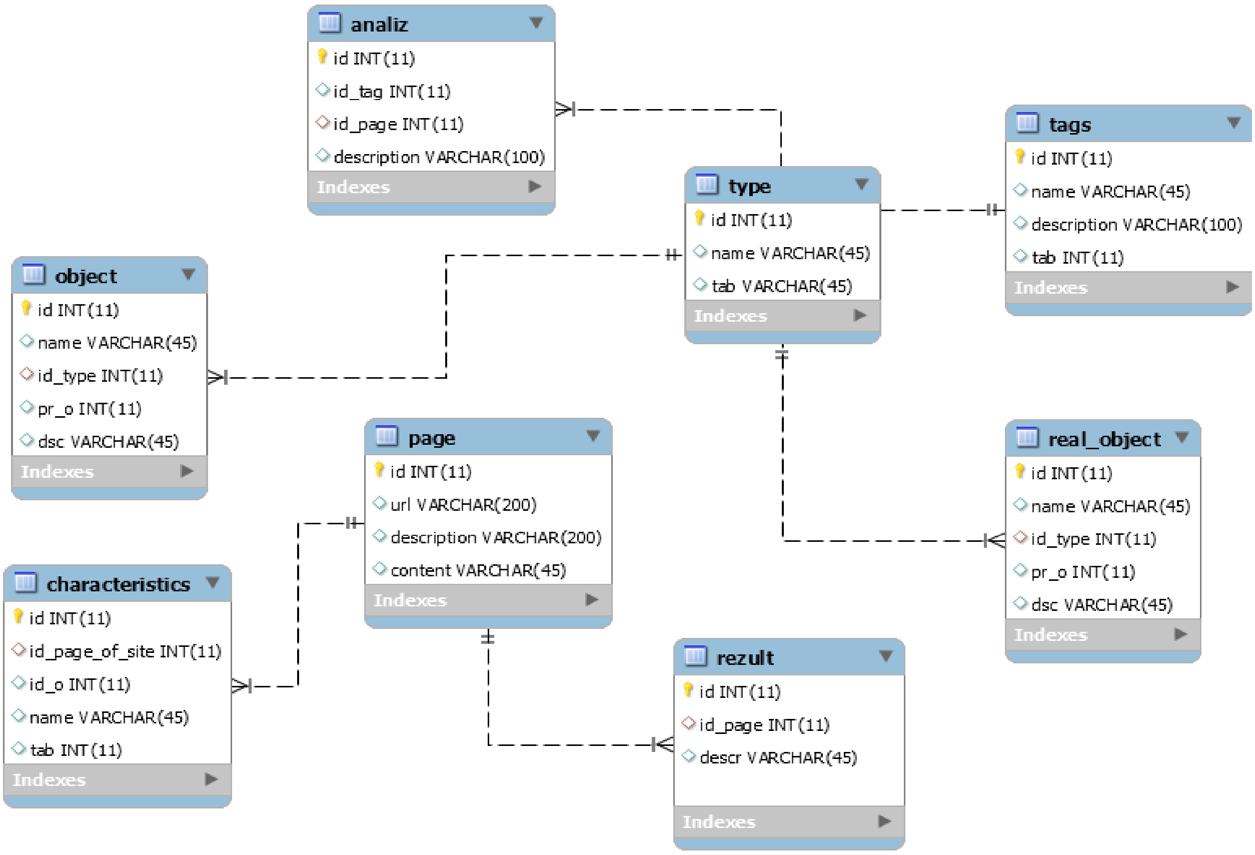


Figure 1 – Structure of the information processing and storage subsystem for the program of detecting irrelevant and inaccurate information on web resources and comparison with the existing database

DOI 10.15588/1607-3274-2018-3-14

If this characteristic coincides with relevant information, then its O_Dsc coincides with the corresponding relevant value. If values of these corresponding characteristics do not coincide, then the information needs to be updated [15, 16].

Based on the analysis of existing systems for websites content analysis, we can formulate the basic requirements for automated algorithms for detection of outdated and incorrect information Fig. 2:

– setting of URL filters, in order to not to parse some extra pages;

– setting of parsing depth;

– high quality content downloading;

– multi stream processing and saving of content in various formats.

After the parsing, generated results must be processed in order to provide the representation of information in such form that is suitable for further use. The exact format depends on how, in future, the collected data will be processed. Quite often, from parsed content, RSS-stream is formed using XML Fig. 3. It is convenient for using the data without rewriting procedure.

Sometimes, the result of parsing is saved into CSV-file, because this text format is very simple for further processing, easily converted to SQL queries and able to work with in Excel. In special cases, it is needed to represent the final data in form of XLS spreadsheets Fig. 4.

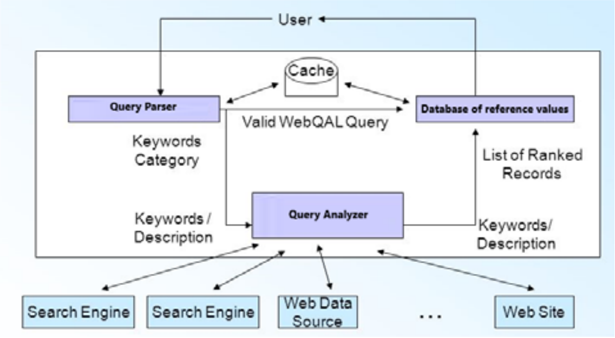


Figure 2 – Architecture of software system for structuring and recognizing of text content of web resources with the machine learning

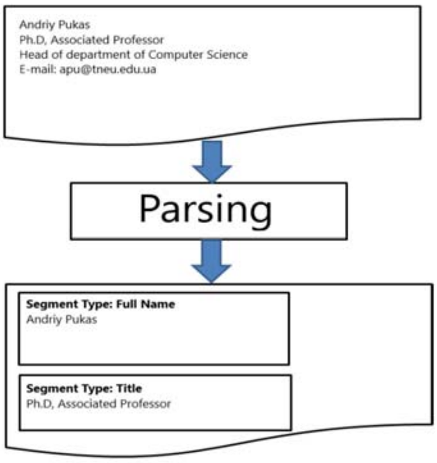


Figure 3 – Schematic illustration of information collecting using parsing, and its storing

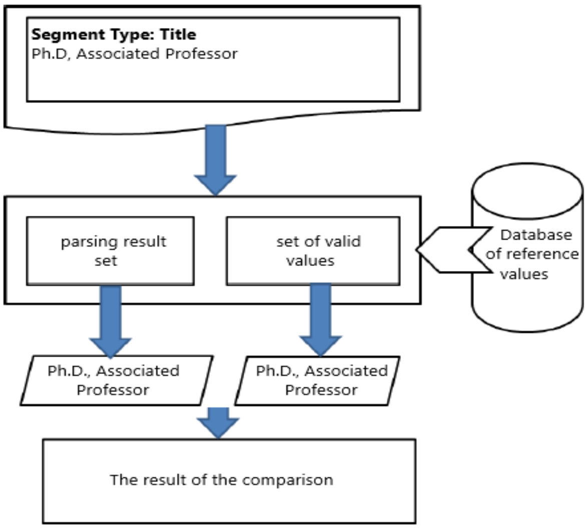


Figure 4 – Schematic illustration of the method for detecting of outdated and incorrect information on web resources and comparison with the existing database

4 EXPERIMENTS

In the process of analysis of text information obtained from web pages, attention must be paid to the process of text normalizing. This will allow to increase the percentage of establishment of equivalence between analyzed and reference.

At the normalization stage, we decrypt some of the abbreviations using corresponding subject dictionaries. In Table 2, for example, the abbreviation "Can. Tech. Sciences" is decrypted as "Candidate of Technical Sciences". Also, we convert a part of the text into a new form. For example, the human initials that are identified in the parsing process, are transformed as it is shown in Table 1. Some phrases are also transformed into more familiar corresponding phrases that are used in the subject area. For example, "The head of the department" is transformed into a "Head of department", as it is shown in Table 2.

Table 1 – Normalization of text information

Input	Output	Type
Can. Tech. Sciences	Candidate of Technical Sciences	Degree
ANDRIY PUKAS	Andriy Pukas	Name
The head of the department	Head of department	Staff

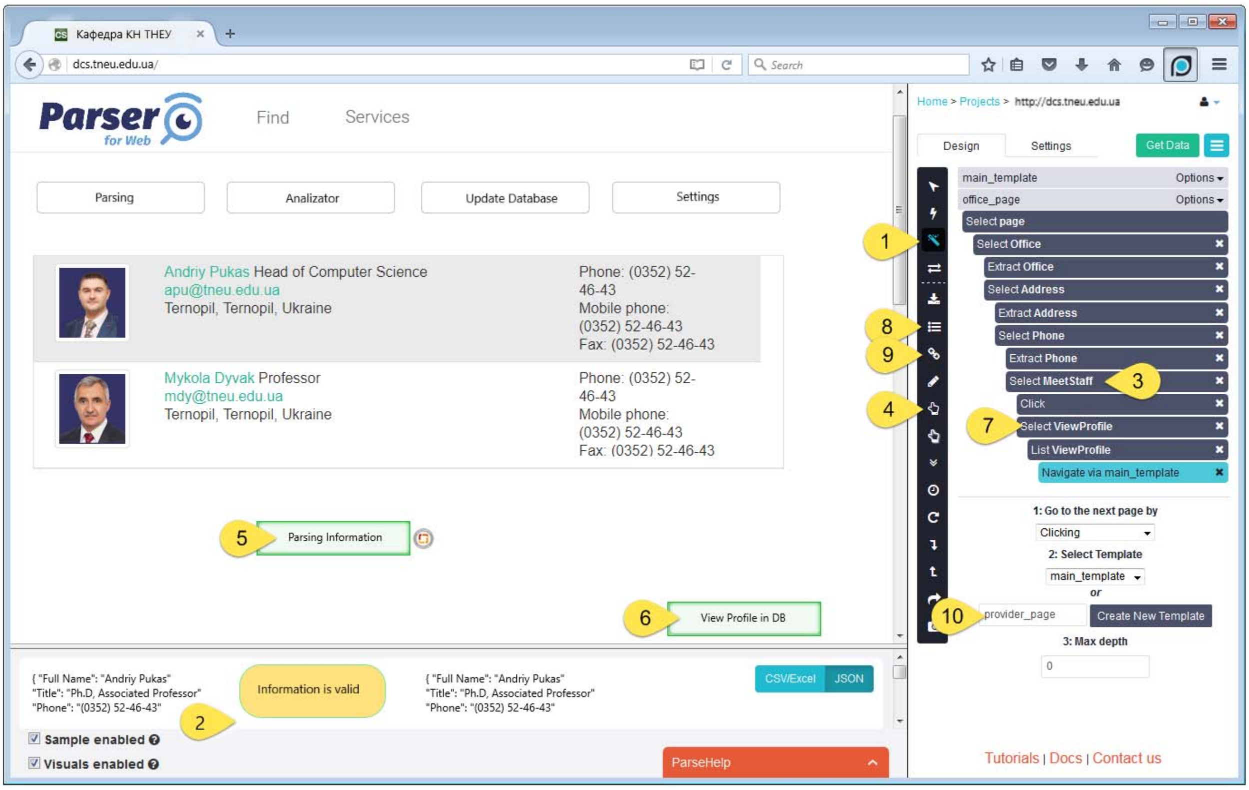
Table 2 – Normalization of text information

Term (Full Form/word)	Abbreviation
Doctor of Philosophy	Ph.D., PhD
Master of Science	M.S., MS, MSc
Bachelor of Computer Application	BCA, B.C.A.

5 RESULTS

Let us consider, in more details, developed system for structuring and recognizing of outdated and incorrect information on the example of the Faculty of Computer Information Technologies of Ternopil National Economic University website (staff information page) Fig. 5.

© Dyvak M. P., Kovbasistyi A. V. Melnyk A. M. Turchyn L. Y., Martsenyuk Y. O., 2018
DOI 10.15588/1607-3274-2018-3-14



Andriy Pukas Head of Computer Science
apu@tneu.edu.ua
Ternopil, Ternopil, Ukraine

Phone: (0352) 52-46-43
Mobile phone: (0352) 52-46-43
Fax: (0352) 52-46-43

Mykola Dyvak Professor
mdy@tneu.edu.ua
Ternopil, Ternopil, Ukraine

Phone: (0352) 52-46-43
Mobile phone: (0352) 52-46-43
Fax: (0352) 52-46-43

```
{ "Full Name": "Andriy Pukas"
"Title": "Ph.D, Associated Professor"
"Phone": "(0352) 52-46-43"
```

Information is valid

```
{ "Full Name": "Andriy Pukas"
"Title": "Ph.D, Associated Professor"
"Phone": "(0352) 52-46-43"
```

☑ Sample enabled
☑ Visuals enabled

1: Go to the next page by
Clicking
2: Select Template
main_template
or
provider_page
3: Max depth
0

Figure 5 – Example of the structuring and recognizing of outdated and incorrect information on the website of educational institution

The lifecycle of the system includes the following stages:

– Choosing a web page or the entire website.

– Checking the validity and relevance of information.

– Selecting object.

– Selecting object attributes.

– Parsing the content of a web page or a website.

– Saving parsing results into database.

– Viewing of selected profile.

– Validation of data with the possibility of updating in a database.

– Comparison of data obtained from the website and data from the organization’s database.

– In case of obtaining of insufficient results, we supplement the rules (parsing, connecting dictionaries).

6 DISCUSSION

Thus, created system for web resources content structuring and recognizing unlike existing ones, contains elements of machine learning. At the same time, the system was built based on parsing algorithms. Similar of them are used in known systems: DataCol, Sjs-parser program system, Content Downloader software system.

The advantage of proposed system is also its cross-platformism. It means that description language of content structure is universal. Realization of the system is adopted for using with different types of relational database management system. Multiple examples of system applying one of which is represented above, showed high performance in recognition, structuring and analysis of content on different types of web resources.

At the same time, developed system has some disadvantages. In particular, the presence of machine learning elements allows to expand the capabilities of the main database which is used for the content analysis, which increases the completeness of content coverage. But, on the other hand, the presence of this element at the same time reduces the reliability of the recognition results and establishment of relevance between analyzed content and basic one.

The development of this work is advisable to be directed into the research of indicators of completeness of content representation in existing databases. And in cases of application of machine learning elements – into determining the relevance of representation of the content structure elements characteristics.

CONCLUSION

The problem of structuring and recognizing of content of web resources with the machine learning elements has been considered.

Scientific novelty of obtained result consists in creation of mathematical tools of system for structuring and recognizing of content of web resources with the machine learning elements. Its main difference is the presence of machine learning components. The proposed system compared with the known ones, due to the above mentioned fact, ensures automated content structuring, recognition of outdated, non-reliable or wrong information. Represented example confirms the effectiveness of the proposed system.

Practical significance of obtained result consists in developing of software that may be used by support services in order to update and correct the content information. The software was develop based on the algorithms that execute the following functions: setting of URL filters, in order to not to

© Dyvak M. P., Kovbasistyi A. V. Melnyk A. M. Turchyn L. Y., Martsenyuk Y. O., 2018
DOI 10.15588/1607-3274-2018-3-14

parse some extra pages; setting of parsing depth; high quality content downloading; multi stream processing and saving of content in various formats.

Prospects for further research are extension of functional possibilities of the systems in the way of use of neural networks elements for machine learning and increasing the quality of web resources content structuring.

REFERENCES

1. Abernethy J., Bartlett P., Rakhlin A., Tewari A. Optimal strategies and minimax lower bounds for online convex games, *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory, COLT 2008, Pittsburgh,* PA, USA, June 22–25, 2008, pp. 1–15. DOI: 10.1007/11776420
2. Aone C., Bennett S. W. Applying machine learning to anaphora resolution, *Connectionist, statistical and symbolic approaches to learning for natural language processing.* Berlin, Springer-Verlag, 1996, pp. 302–314. DOI: 10.1007/3-540-60925-3_55
3. Ayodele T. O. Types of machine learning algorithms in New Advances in Machine Learning. Croatia, Rijeka, InTech, 2010, pp. 19–48. DOI: 10.5772/9385
4. Barber D., Bartlett P., Bousquet O., Mendelson S. Bayesian reasoning and machine learning, *Local rademacher complexities. Annals of Statistics,* 2005, Vol. 33, Issue. 4, pp. 1497–1537. DOI: 10.1145/2636805.2636813
5. Bengio Y. Learning deep architectures for AI *Foundations and Trends in Machine Learning,* 2009,Vol. 2, Issue 1, pp. 1–127. DOI: 10.1561/2200000006
6. Gerbic P., Stacey E. A Purposive approach to content analysis: designing analytical frameworks, *Internet and Higher Education*, 2005, pp. 845–859. DOI: 10.1016/j.iheduc.2004.12.003
7. Types of Machine Learning Algorithms [Electronic resource]. Access mode : http://cdn.intechopen.com/pdfs/10694/InTech-Types_of_machine_learning_algorithms.pdf
8. Harrington P. Machine Learning in Action. New York, Shelter Island, 2012, P. 66–77.
9. Kovbasistyi A., Melnyk A., Dyvak M., Brych V. et al. Method for detection of non-relevant and wrong information based on content analysis of web resources, *XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH).* Lviv, 2017, pp. 154–156. DOI: 10.1109/MEMSTECH.2017.7937555
10. Le Q. V, Ranzato M.-A, Monga R., Devin M. et al. Building high-level features using large scale unsupervised learning [Electronic resource], *International Conference on Machine Learning (ICML),* 26–31 May 2013, Access mode : https://ieeexplore.ieee.org/document/6639343/
11. Mayring P. Qualitative content analysis [Electronic resource], *Forum: Qualitative Social Research.* Access mode : http://217.160.35.246/fqs-texte/2-00/2-00mayring-e.pdf
12. Smola A., Viswanathan S. V. N. Introduction to Machine Learning [Electronic resource] : eBook. Cambridge University Press, 2008, P. 234. Access mode : https://www.kth.se/social/upload/5397442af27654381071d167/chap1.pdf
13. Tools for parsing in the work of an SEO specialist [Electronic resource]. Access mode: https://netpeak.net/ru/blog/instrumenty-dlya-parsinga-v-rabote-seo-spetsialista/.
14. Weare C., Lin W. Y. Content Analysis of the World Wide Web: Opportunities and Challenges, *Social Science Computer Review,* 2002, Vol. 18, P. 272. DOI: 10.1177/089443930001800304
15. Witten Ian H., Eibe Frank, Mark Hal Data Mining. Practical Machine Learning Tools and Techniques [Electronic resource], [3rd Edition.]. San Mateo, Morgan Kaufmann, 2011. Access mode: https://www.elsevier.com/books/data-mining-practical-machine-learning-tools-and-techniques/witten/978-0-12-374856-0
16. Xu Lin, Holger H. Hoos, Kevin Leyton-Brown Hydra Automatically configuring algorithms selection, *In Twenty-Fourth Conference of the Association for the Advancement of Artificial Intelligence*, 2010, pp. 210–216.

Received 25.06.2018.
Accepted 31.07.2018.

УДК 004.774

СИСТЕМА СТРУКТУРУВАННЯ ТА РОЗПІЗНАВАННЯ КОНТЕНТУ ВЕБ-РЕСУРСІВ ІЗ ЕЛЕМЕНТАМИ МАШИННОГО НАВЧАННЯ

Дивак М. П. – д-р техн. наук, професор, декан факультету комп’ютерних інформаційних технологій Тернопільського національного економічного університету, Тернопіль, Україна.

Ковбасістий А. В. – аспірант кафедри комп’ютерних наук Тернопільського національного економічного університету, Тернопіль, Україна.

Мельник А. М. – канд. техн. наук, доцент кафедри комп’ютерних наук Тернопільського національного економічного університету, Тернопіль, Україна.

Турчин Л. Я. – канд. екон. наук, доцент кафедри підприємництва, торгівлі та маркетингу Тернопільського національного економічного університету, Тернопіль, Україна.

Марценюк Є. О. – канд. техн. наук, доцент кафедри комп’ютерних наук Тернопільського національного економічного університету, Тернопіль, Україна.

АНОТАЦІЯ

Актуальність. Наявність великої кількості веб-ресурсів різних організацій вимагає перевірки актуальності та достовірності контенту, зокрема, який стосується характеристик організації, персоналу і т.д. Для цього необхідно розробити систему автоматизованого аналізу контенту. Зазначена задача породжує потребу у розробці методу та програмного забезпечення для структурування та розпізнавання вмісту веб-ресурсів. Існуючі системи парсингу не забезпечують розв’язування зазначеного завдання, оскільки не містять елементів машинного навчання. Об’єктом дослідження є процес автоматизованого аналізу вмісту веб-ресурсів.

Мета роботи – створення системи структурування та розпізнавання вмісту веб-ресурсів з елементами машинного навчання.

Метод. Розглянута система структурування та розпізнавання текстового вмісту веб-ресурсів із елементами машинного навчання. Запропоновані моделі функціонування системи. Розроблено архітектуру для реалізації програмної системи для структурування та розпізнавання текстового вмісту веб-ресурсів. Наведено приклад реалізації моделі розробленої системи для структурування, визнання та виявлення застарілих та невірних відомостей про персонал на веб-ресурсі навчального закладу.

Результати. Розроблений формалізований опис елементів машинного навчання та на його основі програмне забезпечення може використовуватися службою підтримки для оновлення та виправлення контенту веб-ресурсів різних організацій.

Висновки. Розглянута система структурування та розпізнавання вмісту веб-ресурсів із елементами машинного навчання. Пропонована система в порівнянні з відомими, забезпечує автоматичне структурування вмісту, визнання застарілої, недостовірної або неправильної інформації. Представлений приклад структурування та визнання застарілої та некоректної інформації на веб-сайті навчального закладу підтверджує ефективність запропонованої системи.

КЛЮЧОВІ СЛОВА: аналіз контенту, парсинг, машинне навчання.

DOI 10.15588/1607-3274-2018-3-14

УДК 004.774

СИСТЕМА СТРУКТУРИРОВАНИЯ И РАСПОЗНАВАНИЯ КОНТЕНТА ВЕБ-РЕСУРСОВ С ЭЛЕМЕНТАМИ

Дывак Н. П. – д-р техн. наук, профессор, декан факультета компьютерных информационных технологий Тернопольского национального экономического университета, Тернополь, Украина.

Ковбасистый А. В. – аспирант кафедры компьютерных наук Тернопольского национального экономического университета, Тернополь, Украина.

Мельнык А. Н. – канд. техн. наук, доцент кафедры компьютерных наук Тернопольского национального экономического университета, Тернополь, Украина.

Турчын Л. Я. – канд. екон. наук, доцент кафедры предпринимательства, торговли и маркетинга Тернопольского национального экономического университета, Тернополь, Украина.

Марценюк Е. А. – канд. техн. наук, доцент кафедры компьютерных наук Тернопольского национального экономического университета, Тернополь, Украина.

АННОТАЦИЯ

Актуальность. Наличие большого количества веб-ресурсов различных организаций требует проверки актуальности и достоверности контента, в частности, касающийся характеристик организации, персонала и т.д. Для этого необходимо разработать систему автоматизированного анализа контента. Указанная задача порождает потребность в разработке метода и программного обеспечения для структурирования и распознавания содержимого веб-ресурсов. Существующие системы парсинга не обеспечивают решения указанной задачи, поскольку не содержат элементов машинного обучения. Объектом исследования является процесс автоматизированного анализа содержимого веб-ресурсов.

Цель работы – создание системы структурирования и распознавания содержимого веб-ресурсов.

Метод. Рассмотрена система структурирования и распознавания текстового содержимого веб-ресурсов с элементами машинного обучения. Предложены модели функционирования системы. Разработана архитектура для реализации программной системы для структурирования и распознавания текстового содержимого веб-ресурсов. Приведен пример реализации модели разработанной системы для структурирования, выявления устаревших и неверных сведений о персонале на веб-ресурсе учебного заведения.

Результаты. Разработанное программное обеспечение может использоваться службой поддержки для обновления и исправления информационного содержания.

Выводы. Рассмотрена система структурирования и распознавания содержимого веб-ресурсов с элементами машинного обучения. Предлагаемая система по сравнению с известными, обеспечивает автоматическое структурирование содержания, выявления устаревшей, недостоверной или неправильной информации. Представлен пример структурирования и признание устаревшей и некорректной информации на сайте учебного заведения подтверждает эффективность предложенной системы.

КЛЮЧЕВЫЕ СЛОВА: анализ контента, парсинг, машинное обучение.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Optimal strategies and minimax lower bounds for online convex games / J. Abernethy, P. Bartlett, A. Rakhlin, A. Tewari // Proceedings of the Nineteenth Annual Conference on Computational Learning Theory, COLT 2008, Pittsburgh, PA, USA, June 22-25. – 2008. – P. 1–15. DOI: 10.1007/11776420
2. Aone C. Applying machine learning to anaphora resolution / C. Aone, S. W. Bennett // Connectionist, statistical and symbolic approaches to learning for natural language processing. – Berlin : Springer-Verlag, 1996. – P. 302–314. DOI: 10.1007/3-540-60925-3_55
3. Ayodele T. O. Types of machine learning algorithms in New Advances in Machine Learning / T. O. Ayodele. – Croatia, Rijeka : InTech, 2010. – P. 19–48. DOI: 10.5772/9385
4. Barber, D. Bayesian reasoning and machine learning / D. Barber, P. Bartlett, O. Bousquet, S. Mendelson // Local rademacher complexities. Annals of Statistics. – 2005. – Vol. 33, Issue 4. – P. 1497–1537. DOI: 10.1145/2636805.2636813
5. Bengio, Y. Learning deep architectures for AI / Y. Bengio // Foundations and Trends in Machine Learning. – 2009. – Vol. 2, Issue 1. – P. 1–127. DOI: 10.1561/2200000006
6. Gerbic P. A Purposive approach to content analysis: designing analytical frameworks [Text] / P. Gerbic, E. Stacey // Internet and Higher Education. – 2005. – P. 845–859. DOI: 10.1016/j.iheduc.2004.12.003
7. Types of Machine Learning Algorithms [Electronic resource]. – Access mode : http://cdn.intechopen.com/pdfs/10694/InTech-Types_of_machine_learning_algorithms.pdf
8. Harrington P. Machine Learning in Action / P. Harrington. – New York : Shelter Island, 2012. – P. 66–77.
9. Method for detection of non-relevant and wrong information based on content analysis of web resources / [A. Kovbasistyi, A. Melnyk, M. Dyvak et al.] // XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH). – Lviv, 2017. – P. 154–156. DOI: 10.1109/MEMSTECH.2017.7937555
10. Building high-level features using large scale unsupervised learning [Electronic resource] / [Q. V Le, M.-A Ranzato, R. Monga et al.] // International Conference on Machine Learning (ICML). – 26–31 May 2013. – Access mode : https://ieeexplore.ieee.org/document/6639343/
11. Mayring P. Qualitative content analysis [Electronic resource] / P. Mayring // Forum: Qualitative Social Research. – Access mode : http://217.160.35.246/fqs-texte/2-00/2-00mayring-e.pdf
12. Smola A. Introduction to Machine Learning [Electronic resource] : eBook / Alex Smola and S. V. N. Viswanathan. – Cambridge University Press, 2008. – P. 234. Access mode : https://www.kth.se/social/upload/5397442af27654381071d167/chap1.pdf
13. Tools for parsing in the work of an SEO specialist [Electronic resource]. – Access mode: https://netpeak.net/ru/blog/instrumenty-dlya-parsinga-v-rabote-seo-spetsialista/.
14. Weare C. Content Analysis of the World Wide Web: Opportunities and Challenges / C. Weare, W. Y. Lin // Social Science Computer Review. – 2002. – Vol. 18. – P. 272. DOI: 10.1177/089443930001800304
15. Witten, Ian H. Data Mining. Practical Machine Learning Tools and Techniques [Electronic resource] / Ian H. Witten, Eibe Frank, Mark Hal – [3rd Edition.]. – San Mateo : Morgan Kaufmann, 2011. – Access mode: https://www.elsevier.com/books/data-mining-practical-machine-learning-tools-and-techniques/witten/978-0-12-374856-0
16. Xu Lin Hydra Automatically configuring algorithms selection / Lin Xu Hydra, Holger H. Hoos, Kevin Leyton-Brown // In Twenty-Fourth Conference of the Association for the Advancement of Artificial Intelligence. – 2010. – P. 210–216.

DOI 10.15588/1607-3274-2018-3-14