МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Кваліфікаційна наукова
праця на правах рукопису

КОВБАСІСТИЙ АНДРІЙ ВІКТОРОВИЧ

УДК: 519.7:004

ДИСЕРТАЦІЯ

Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості

121 – Інженерія програмного забезпечення
12 – Інформаційні технології

Подається на здобуття ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____Ковбасістий А. В.
Науковий керівник Дивак Микола Петрович доктор технічних наук, професор.

Тернопіль 2021

АНОТАЦІЯ

Ковбасістий А. В. Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості - Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття ступеня доктора філософії за спеціальністю 121 «Інженерія програмного забезпечення» - Західноукраїнський національний університет, Тернопіль, 2020.

Підготовка здійснювалась на кафедрі комп'ютерних наук Західноукраїнського національного університету міністерства освіти і науки України.

Дисертаційна робота присвячена розв'язуванню актуального науково-технічного завдання розробки математичного та програмного забезпечення для автоматизованого структурування та розпізнавання контенту із використанням засобів машинного навчання, яке у сукупності забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів.

У першому розділі проведено аналіз методів, засобів та технологій організації інформаційних веб-ресурсів.

Розвиток інформаційних технологій забезпечує розвиток суспільства, переоцінювання його пріоритетів, особливо модернізацію існуючих галузей економіки та створення і стійкий розвиток нових. Цифровізація усіх сфер діяльності людини сьогодні є життєвою необхідністю. Одним із засобів розв'язування потреб суспільства є інформаційні веб ресурси. Їхня сфера застосування - від сайтів візитівок (статичних) організацій до інтелектуальних веб-ресурсів для надання різноманітних послуг.

Разом з тим, проведений аналіз показав, що ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту а також наявність в контенті застарілої та неактуальної інформації. Тому її виявлення та оновлення є актуальною задачею, розв'язування якої лежить у сфері розробки засобів автоматизованого структурування та розпізнавання контенту.

Далі у цьому розділі проаналізовано відомі системи контент аналізу із використанням засобів штучного інтелекту. Проведено порівняння концепцій організації технологій: Веб 1.0, Веб 2.0, Веб 3.0 та Веб 4.0., а також можливостей використання в процедурах контент аналізу. Проаналізовано алгоритми та засоби організації машинного навчання з метою структурування та розпізнавання контенту, а також методи видобування даних з веб-ресурсів. У цьому розділі також обґрунтовується показник ефективності інформаційних веб-ресурсів, який стосується витрат часу на обробку запитів. Для використання цього показника запропоновано розробити його математичне представлення і провести комп'ютерне моделювання та верифікацію запропонованого математичного опису з метою дослідження впливу структурованості контенту на ефективність веб-ресурсу. Такий підхід забезпечить не тільки встановлення цього впливу для конкретного веб-ресурсу, але і забезпечить моделювання та оцінку структурованості контенту для розв'язування задач його оптимізації, аналізу та перевірки на цілісність, достовірність та актуальність.

У завершальній частині розділу здійснено постановку завдань дисертаційного дослідження.

У другому розділі дисертаційної роботи запропоновано та обґрунтовано новий метод моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту, який дає можливість врахувати характерні для нелінійних систем особливості, такі як залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристик. Отримані на основі запропонованого методу математичні моделі забезпечують вирішення цілого ряду практичних завдань, що виникають при розробці і впровадженні веб-систем: налаштування параметрів продуктивності, аналіз впливу структурованості контенту, оцінка ефективності функціонування під навантаженням. Для побудови математичної моделі ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості, розв'язано задачу ідентифікації параметрів цієї моделі. Своєю чергою, для реалізації процедури ідентифікації параметрів моделі запропоновано

методику організації експериментальних досліджень веб-ресурсу, а для отримання даних про потрібний контент та збір статистики на веб-ресурсі застосовано Google API. Збір статистики часу виконання запитів, здійснювався за допомогою Core Reporting API.

В завершальній частині розділу наведено приклад ідентифікації математичної моделі ефективності функціонування інформаційних веб-ресурсів для конкретного інформаційного веб-ресурсу.

У третьому розділі розглянуто методи та засоби виявлення некоректної та неактуальної інформації на основі контент-аналізу веб-ресурсів.

Сформульовано основні вимоги до засобів автоматичного аналізу контенту, основним завданням яких є виявлення застарілої та некоректної інформації. Встановлено, що засоби повинні бути налаштовані на типові структури веб-сторінок, а також на відомі структури опису веб-сторінки мовою розмітки HTML. Також показано основні складнощі щодо виявлення тегів із різними іменами та порівняння результатів із відомими даними, записаними в базах даних організацій чи виявлені на інших веб-ресурсах

Запропоновано та обґрунтовано підхід до побудови контент аналізу з автоматичним виконанням функцій. Основна ідея побудови системи полягає в перетворенні інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, існуючими в організаціях або виявлених на інших веб-ресурсах з їх вибором за принципами «мажоритарного голосування» та з подальшим порівнянням. Наведено приклади використання запропонованого підходу автоматизованого пошуку застарілої та некоректної інформації на ряді сайтів вищих навчальних закладів, де було отримано доступ до відповідних баз даних.

Запропоновано та обґрунтовано новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання. Запропонований метод у порівняні з відомими, за рахунок уведення формального опису контенту та компонентів машинного навчання, забезпечує автоматизоване структурування контенту, виявлення в структурі контенту недостовірної, неточної чи неактуальної інформації.

У четвертому розділі наведено результати розробки прикладної програмної системи для підтримки функціонування інформаційних веб-ресурсів. Обґрунтовано та сформульовано вимоги до прикладної системи. Розроблено архітектуру програмної системи та структуру бази даних структурування контенту, що виконує функції автоматизованого пошуку застарілої та некоректної інформації. Спроектовано діаграму класів та діаграму компонентів інтелектуальної системи для аналізу структурованості контенту веб-ресурсів. Наведено результати розробки програмного інтерфейсу.

У завершальній частині цього розділу наведено результати застосування прикладної програмної системи для підтримки функціонування інформаційних веб-ресурсів, зокрема, для підтримки функціонування інформаційного веб-ресурсу Західноукраїнського національного університету та дослідження його ефективності, а також для розпізнавання та структурування контенту на прикладі інформаційного веб-ресурсу Центру надання адміністративних послуг Тернопільської міської ради. Результатами апробації підтверджено достовірність положень щодо розробки методу структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання. Також підтверджено змістовність гіпотез, на яких ґрунтується запропонований підхід до побудови контент аналізу.

Практична цінність роботи полягає у створенні прикладної програмної системи, яка забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів за рахунок застосування процедур структурування, розпізнавання та порівняння текстового контенту веб-ресурсів, виявлення в структурі контенту недостовірної, неточної чи неактуальної інформації.

Розроблену прикладну програмну систему апробовано для виявлення некоректної та недостовірної інформації на інформаційному веб-ресурсі Західноукраїнському національного університеті. Застосування розробленого програмного забезпечення та математичної моделі показника ефективності інформаційного веб-ресурсу дало можливість сформувати вимоги для служби SEO веб-ресурсу Центр надання адміністративних послуг Тернопільської міської ради та ТОВ «ТЕРНОГАЗ», зокрема щодо налаштування параметрів

продуктивності, зменшення впливу структурованості контенту на продуктивність ресурсу та оцінити його ефективність функціонування під навантаженням.

ПЕРЕЛІК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковано основні наукові результати дисертації:

1. Dyvak M. P., Melnyk A. M., Kovbasistyi A.V., Turchyn L. Y., Martsenyuk Y. O. System for web resources content structuring and recognizing with the machine learning elements // Radio Electronics, Computer Science, Control. - Zaporizhzhia, No 3(46). - 2018. - P. 128-134. (Web of Science)

2. Dyvak M., Kovbasistyi A., Stakhiv P., Lipiński P. Generalized structure of the algorithm for automated detection of non relevant and wrong information on web resources // Journal of Applied Computer Science. - 2017. - P. 23-37.

3. Дивак М.П., Мельник А. М., Ковбасістий А. В., Папа О. А. Підхід до математичного моделювання ефективності WEB-ресурсів // Оптико-електронні інформаційно-енергетичні технології, Міжнародний науково-технічний журнал. - №2 (38). - 2019. - С. 29-38.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

4. Дивак М. П., Ковбасістий А. В. Особливості побудови методу виявлення некоректної та застарілої інформації. Сучасні комп'ютерні інформаційні технології. Матеріали VI Всеукраїнської школи-семінару молодих вчених і студентів (АСІТ’2016). - Тернопіль: ТНЕУ. - 2016. - С. 119-120.

5. Melnyk A., Dyvak M., Kovbasistyi A., Brych V., Spivak I. Method for Detection of Non-Relevant and Wrong Information Based on Content Analysis of Web Resources. Proc. of the 2017 XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH). - 20-24 April, 2017. - Polyana-Svalyava (Zakarpattya), UKRAINE. – P. 94-96. (IEEE Xplore Digital Library, DOI: 10.1109/MEMTECH.2017.7937555 – P. 94-96.) (Scopus).

6. Dyvak M., Melnyk A., Kovbasistyi A., Shcherbiak I., Huhul O. Recognition of relevance of web resource content based on analysis of semantic components Proc. of the IX International Conference on Advanced Computer

Information Technologies (ACIT'2019).– Ceske Budejovice, Czech Republic. - 2019. - P. 297-302 (IEEE Xplore Digital Library, DOI: 10.1109/ACITT.2019.8779897) (Scopus).

7. Dyvak M., Melnyk A., Kovbasistyi A., Shevchuk R., Huhul O., Tymchyshyn V. Mathematical Modeling of the Estimation Process of Functioning Efficiency Level of Information Web-Resources. - Proc. of the 10th International Conference on Advanced Computer Information Technologies (ACIT’2020). – Deggendorf, GERMANY. - 16-18 September 2020. - P. 492-496 (IEEE Xplore Digital Library, DOI: 10.1109/ACIT49673.2020.9208846) (Scopus).

ANNOTATION

Kovbasisty A.V. Mathematical and software support for the operation of web information resources in different configurations – Published as a manuscript.

The Ph.D thesis in “Software engineering” (specialty 121) - West Ukrainian National University, Ternopil, 2020.

The graduate Ph. D. program was performed at the Department of Computer Science of the West Ukrainian National University of the Ministry of Education and Science of Ukraine.

The thesis is devoted to solving the current scientific and technical problem of developing mathematical and software support for automated structuring and recognition of content using machine learning tools, which increase the efficiency of information web resources.

The first section analyzes the methods, tools, and technologies of organizing information web resources.

The advancement of information technology ensures the development of society, reassessment of its priorities, especially the modernization of existing sectors of the economy, and the creation and sustainable development of new ones. The digitalization of all spheres of human activity is a vital necessity. One way to address society’s needs is to provide informational web resources. Their scope ranges from business card sites (static) organizations to intelligent web resources to provide a variety of services.

However, the analysis showed that the effectiveness of most web resources is low because of the weak structure of the content and outdated information. Therefore, its detection and updating is an urgent task, the solution of which lies in the development of automated structuring and content recognition tools.

Further on the part analyzes known AI content analysis systems. A comparison was made between Web 1.0, Web 2.0, Web 3.0, and Web 4.0. as well as the possibilities of using in Content Analysis procedures. Algorithms and means of organizing machine learning for structuring and recognizing content, as well as methods of extracting data from web resources are analyzed. The part also substantiates the performance indicator of information web resources relating to the span of

processing requests. It is proposed to develop a mathematical representation and conduct both computer modeling and verification of the mathematical description with the indicator. Within this, the impact of content structure on the effectiveness of the web resource has been studied. The approach will establish the impact for a particular web resource and provide modeling and evaluation of the content to solve such problems as optimization, analysis, and verification for integrity, reliability, and relevance.

The research tasks are set in the final part.

The second part of the thesis proposes and substantiates a new method of modeling the indicator for assessing the effectiveness of information web resources depending on the structure of content, which allows taking into account the characteristics of nonlinear systems, such as perturbation dependence on current values, cyclical processes, and abrupt changes. The mathematical models obtained based on the proposed method provide a solution to several practical problems that arise in the development and implementation of web systems: setting performance parameters, analyzing the impact of content structure, assessing the effectiveness of functioning under load. To build a mathematical model of the efficiency of information web resources in terms of their different structure, the problem of identifying the parameters of this model is solved. To implement the procedure for identifying the parameters of the model a method of organizing experimental studies of the web resource is proposed, and Google API is used to obtain data on the required content and collect statistics on the web resource. Collection of statistics on inquiries processing span was carried out utilizing Core Reporting API.

The final part of the section provides an example of identifying a mathematical model of the effectiveness of information web resources for a particular information web resource.

The third section discusses the methods and means of detecting incorrect and outdated information based on content analysis of web resources.

The basic requirements to the means of automatic content analysis are formulated. The main task of them is to identify outdated and incorrect information. It is established that the tools must be adapted to the typical structures of web pages, as well as to the

known structures of the description of the web page in HTML markup language. It also shows the main difficulties in identifying tags with different names and comparing the results with known data recorded in the databases of organizations or found on other web resources.

The approach to the content analysis with an automatic performance of functions is offered and proved. The main idea of building a system is to convert information from web resources into a format suitable for comparison with databases existing in organizations or found on other web resources with their choice on the principles of "majority voting" and subsequent comparison. Examples of using the proposed approach of automated search of outdated and incorrect information on several sites of higher education institutions, where access to the relevant databases was obtained, are given.

A new method of structuring, recognizing, and comparing the textual content of web resources with the elements of machine learning is proposed and substantiated. The proposed method, thanks to introduction of a formal description of the content and components of machine learning, provides automated structuring of content, detection of unreliable, inaccurate, or irrelevant information in the content structure.

The fourth section presents the results of the development of an application software system to support the functioning of information web resources. The requirements for the application system are substantiated and formulated. The architecture and structure of the structured content database for the software system that performs the functions of an automated search for outdated and incorrect information has been developed. The proposed architecture and generalized algorithms for implementing its components have been tested on several examples from web resources of educational institutions, which confirms the reliability of the results of developing a method of structuring, recognizing, and comparing textual content of web resources with elements of machine learning to build content analysis.

The class diagram and the diagram of the components of the intelligent system for the analysis of the structure of the content of web resources are designed. The results of software interface development are given.

The final part of this section presents the results of the application of the software system to support the functioning of information web resources, in particular, to the functioning of the information web resource of West Ukrainian National University and study its effectiveness, as well as to recognize and structure the content of Administrative Services Center of Ternopil City Council.

The practical significance of the work is to create an application software system that improves the efficiency of information web resources through structuring, recognizing, and comparing textual content of web resources, identifying inaccurate or irrelevant information in the content structure.

The developed application software system has been tested to detect incorrect and unreliable information on the web resource of West Ukrainian National University. The application of the developed software and mathematical model of the efficiency indicator of the information web resource made it possible to form requirements for SEO web resource of Ternopil City Council Administrative Services Center, in particular, to adjust performance parameters, reduce the impact of content structure on resource performance and evaluate its effectiveness.

PUBLICATION LIST BY THE DISSERTATION SUBJECT

Publications in which the main scientific results of the dissertation has been published:

1. Dyvak M. P., Melnyk A. M., Kovbasistyi A.V., Turchyn L. Y., Martsenyuk Y. O. System for web resources content structuring and recognizing with the machine learning elements // Radio Electronics, Computer Science, Control. - Zaporizhzhia, No 3(46). - 2018. - P. 128-134. (Web of Science)

2. Dyvak M., Kovbasistyi A., Stakhiv P., Lipiński P. Generalized structure of the algorithm for automated detection of non relevant and wrong information on web resources // Journal of Applied Computer Science. - 2017. - P. 23-37.

3. Dyvak, M. P., Melnyk, A. M., Kovbasistyi, A. V., Papa, O. A. (2020). Approach to Mathematical Modelling of Web Resources Efficiency. Optoelectronic information and energy technologies, 38(2), 29–37. https://doi.org/10.31649/1681-7893-2019-38-2-29-37

Publications certifying the approbation of the dissertation materials:

4. Dyvak, M. P., Kovbasistyi, A. V. Features of building a method for detecting non-relevant and wrong information. Modern computer information technologies. Proceedings of the VI All-Ukrainian school-seminar of young scientists and students (ACIT’2016). - Ternopil: TNEU. - 2016. - C. 119-120.

5. Melnyk A., Dyvak M., Kovbasistyi A., Brych V., Spivak I. Method for Detection of Non-Relevant and Wrong Information Based on Content Analysis of Web Resources. Proc. of the 2017 XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH). - 20-24 April, 2017. - Polyana-Svalyava (Zakarpattya), UKRAINE. – P. 94-96. (IEEE Xplore Digital Library, DOI: 10.1109/MEMTECH.2017.7937555 – P. 94-96.) (Scopus).

6. Dyvak M., Melnyk A., Kovbasistyi A., Shcherbiak I., Huhul O. Recognition of relevance of web resource content based on analysis of semantic

components Proc. of the IX International Conference on Advanced Computer Information Technologies (ACIT'2019).– Ceske Budejovice, Czech Republic. - 2019. - P. 297-302 (IEEE Xplore Digital Library, DOI: 10.1109/ACITT.2019.8779897) (Scopus).

7. Dyvak M., Melnyk A., Kovbasistyi A., Shevchuk R., Huhul O., Tymchyshyn V. Mathematical Modeling of the Estimation Process of Functioning Efficiency Level of Information Web-Resources. - Proc. of the 10th International Conference on Advanced Computer Information Technologies (ACIT’2020). – Deggendorf, GERMANY. - 16-18 September 2020. - P. 492-496 (IEEE Xplore Digital Library, DOI: 10.1109/ACIT49673.2020.9208846) (Scopus).

Зміст

АНОТАЦІЯ 2

ВСТУП 17

РОЗДІЛ 1 АНАЛІЗ МЕТОДІВ, ЗАСОБІВ ТА ТЕХНОЛОГІЙ ОРГАНІЗАЦІЇ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ 24

1.1. Особливості структурування та організації веб-ресурсів 25

1.2. Огляд відомих систем контент аналізу 32

1.2.1. Веб-технології як основа побудови систем аналізу контенту з елементами машинного навчання 32

1.2.2. Огляд відомих спеціалізованих систем контент-аналізу 38

1.3. Обґрунтування показника ефективності функціонування інформаційних веб ресурсів 41

1.4. Постановка завдань дисертаційного дослідження 49

Висновки до першого розділу 53

РОЗДІЛ 2 МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ЕФЕКТИВНОСТІ ФУНКЦІОНУВАННЯ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ В УМОВАХ РІЗНОЇ ЇХ СТРУКТУРОВАНОСТІ 55

2.1. Показник ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості 56

2.2. Математична модель ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості 59

2.3. Методика отримання експериментальних даних 63

2.4. Приклад ідентифікації математичної моделі ефективності функціонування інформаційних веб-ресурсів 69

Висновки до другого розділу 77

РОЗДІЛ 3 МЕТОДИ ТА ЗАСОБИ ВИЯВЛЕННЯ НЕКОРЕКТНОЇ ТА НЕАКТУАЛЬНОЇ ІНФОРМАЦІЇ НА ОСНОВІ КОНТЕНТ-АНАЛІЗУ ВЕБ-РЕСУРСІВ 78

3.1. Методи та засоби виявлення некоректної та неактуальної інформації на основі існуючих баз даних контенту 78

3.2. Метод та засоби виявлення некоректної та неактуальної інформації із застосуванням принципу “мажоритарного голосування” 83

3.3. Метод виявлення некоректної та недостовірного контенту із використанням засобів машинного навчання 89

3.4. Архітектура програмної системи для реалізації методу виявлення некоректного та недостовірного контенту 95

Висновки до третього розділу 99

РОЗДІЛ 4. ПРИКЛАДНА ПРОГРАМНА СИСТЕМА ДЛЯ ПІДТРИМКИ ФУНКЦІОНУВАННЯ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ 100

4.1. Функціональні можливості 100

4.2. Особливості організації інтерфейсу та функціонування 109

4.3. Застосування прикладної програмної системи для підтримки функціонування інформаційних веб-ресурсів. 117

4.3.1 Застосування прикладної програмної системи для підтримки функціонування інформаційного веб-ресурсу Західноукраїнського національного університету та дослідження його ефективності 117

4.3.2. Експериментальні дослідження методу моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів на прикладі Центру надання адміністративних послуг Тернопільської міської ради та ТОВ «Терногаз» 119

Висновки до четвертого розділу 121

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 125

ДОДАТКИ 135

ВСТУП

Актуальність теми.

Розвиток інформаційних технологій призводить до швидкої зміни пріоритетів суспільства, модернізації існуючих галузей економіки та створення і становлення нових. Цифровізація усіх сфер діяльності людини сьогодні є не популярним трендом, а життєвою необхідністю. Одним із засобів розв'язування потреб суспільства є інформаційні веб-ресурси. Їхня сфера застосування: від сайтів візитівок (статичних) організацій до інтелектуальних веб-ресурсів для надання різноманітних послуг. Разом з тим, ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту а також наявність у ньому застарілої та неактуальної інформації. Тому її виявлення та оновлення є актуальним завданням, вирішення якого лежить у сфері розробки засобів автоматизованого структурування та розпізнавання контенту.

Задачі автоматизованого розпізнавання контенту та побудови систем контент-аналізу із використанням засобів штучного інтелекту розглянуто в працях таких вітчизняних та зарубіжних вчених: B Efron, RJ Tibshirani, R Kohavi.

Проте проблема структурування контенту, що є складовою для розв'язування вищезазначеного завдання, залишається актуальною в силу різноманіття веб-ресурсів та мов їх створення.

Питаннями створення та підвищення ефективності функціонування інформаційних веб-ресурсів у рамках технології веб.3.0 займалися як вітчизняні, так і зарубіжні вчені, серед яких особливо виділяються праці А. М. Пелещишин, В. В. Пасічник, J.Kleinberg, S.Brin, L.Page.

Разом з тим, сьогодні вже актуальним є створення технологій веб.4.0, особливостями яких є широке застосування автоматичних систем контент аналізу із використанням засобів штучного інтелекту для структурування та розпізнавання контенту, а також ґрунтовним підвищенням ефективності інформаційних веб-ресурсів.

У межах сучасних технологій веб.3.0 та веб.4.0 особлива увага приділяється дослідженню та підвищенню ефективності створюваних й існуючих веб-ресурсів. Проблеми, пов’язані із дослідженням ефективності інформаційних веб-ресурсів, розглянуто у працях R.Kosala, R.Bayeza-Yates, G.Piatetsky-Shapiro, J.Barker.

Разом з тим, розглянуті показники ефективності мають ряд недоліків, більшість з яких пов’язана із слабким математичним обґрунтуванням та формальним представленням.

Виходячи із вищезазначеного, актуальним є вирішення науково-технічного завдання щодо розробки математичного та програмного забезпечення для автоматизованого структурування і розпізнавання контенту із використанням засобів машинного навчання, яке у сукупності забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів.

Створення зазначених засобів дає можливості для розробки цілого ряду прикладних програмних систем для виявлення неактуальної та некоректної інформації на інформаційних веб-ресурсах.

Мета і завдання дослідження. Метою дисертаційного дослідження є підвищення ефективності функціонування інформаційних веб-ресурсів в умовах різної структурованості контенту за рахунок упровадження методів його автоматизованого структурування і розпізнавання, виявлення некоректної та застарілої інформації із використанням засобів машинного навчання.

Для досягнення цієї мети необхідно виконати такі завдання:

- проаналізувати існуючі методи, засоби та технології організації інформаційних веб-ресурсів та проблеми підтримки їх функціонування;
- обґрунтувати базовий показник оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту та розробити метод його налаштування і моделювання;
- розробити метод структурування, розпізнавання і порівняння текстового контенту веб-ресурсів із елементами машинного навчання та з використанням автоматизованого структурування контенту, виявлення недостовірної, неточної чи неактуальної інформації;

– удосконалити методи автоматизованого контент-аналізу, пошуку застарілої та некоректної інформації з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, створеними з метою виявлення подібного контенту на інших веб-ресурсах.

– розробити архітектури прикладних програмних систем, що виконують функції автоматизованого пошуку застарілої та некоректної інформації за рахунок наявності модулів розпізнавання та порівняння текстового контенту;

– провести апробацію запропонованих методів та архітектур прикладних програмних систем при вирішенні ряду прикладних задач щодо контент-аналізу, пошуку застарілої та некоректної інформації інформаційних веб-ресурсів;

Об'єкт дослідження – процеси контент-аналізу на інформаційних веб-ресурсах.

Предмет дослідження – математичне, програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості.

Методи дослідження.

Для розробки методу моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту використано стохастичний підхід та методи параметричної ідентифікації моделей нелінійних систем. Для розробки методів структурування, розпізнавання та порівняння текстового контенту веб-ресурсів – методи семантичного опису та аналізу контенту із елементами машинного навчання. Розвиток й удосконалення архітектур створеного програмного забезпечення, здійснено із використанням методів обєктно-орієнтованого аналізу та проектування. Для проектування та опису програмного забезпечення застосовано середовище UML -2.0

Наукова новизна отриманих результатів.

У межах дисертаційної роботи *вперше:*

- запропоновано та обґрунтовано метод моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту, який, на відміну від існуючих, дає можливість врахувати характерні для нелінійних систем особливості, такі як: залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристик. Отримані на основі запропонованого методу математичні моделі забезпечують вирішення багатьох практичних завдань, що виникають при розробці і впровадженні веб-систем: налаштування параметрів продуктивності, аналіз впливу структурованості контенту, оцінка ефективності функціонування під навантаженням;
- запропоновано й обґрунтовано новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання, що за рахунок уведення формального опису контенту та компонентів машинного навчання забезпечує автоматизоване структурування контенту, виявлення у його структурі недостовірної, неточної чи неактуальної інформації.

Набули подальшого розвитку:

- методи автоматизованого контент-аналізу з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, що відрізняються створенням цих баз даних для виявлення подібного контенту на інших веб-ресурсах з їх вибором за принципами «мажоритарного голосування» та з подальшим порівнянням, що забезпечує підвищення ефективності функціонування веб-ресурсу, та виконання функцій автоматизованого пошуку застарілої і некоректної інформації;
- архітектури прикладних програмних систем, що виконують функції автоматизованого пошуку застарілої та некоректної інформації, які містять модулі розпізнавання і порівняння текстового контенту веб-ресурсів із елементами машинного навчання, що, у свою чергу, забезпечує підвищення ефективності функціонування прикладних програмних систем категорії інформаційні веб-ресурси.

Практичне значення отриманих результатів полягає у створенні прикладної програмної системи, яка забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів за рахунок застосування процедур структурування, розпізнавання і порівняння текстового контенту веб-ресурсів, виявлення у його структурі недостовірної, неточної чи неактуальної інформації.

Розроблену прикладну програмну систему апробовано для виявлення некоректної та недостовірної інформації на інформаційному веб-ресурсі Західноукраїнського національного університету. Застосування розробленого програмного забезпечення та математичної моделі показника ефективності інформаційного веб-ресурсу дало можливість сформувати вимоги для SEO веб-ресурсу Центр надання адміністративних послуг Тернопільської міської ради, зокрема щодо налаштування параметрів продуктивності, зменшення впливу структурованості контенту на продуктивність ресурсу й оцінити його ефективність функціонування під навантаженням.

Теоретичні та прикладні результати дисертаційної роботи використано:

- у Західноукраїнському національному університеті при структуруванні, розпізнаванні та порівнянні текстового контенту веб-ресурсів із елементами машинного навчання та із використанням автоматизованого структурування контенту, виявлення недостовірної, неточної чи неактуальної інформації (акт про впровадження результатів дисертаційної роботи від 11 листопада 2020 р.);
- у центрі надання адміністративних послуг Тернопільської міської ради (акт про впровадження результатів дисертаційної роботи від 04 листопада 2020 р.);
- при виконанні госпдоговірної науково-дослідної роботи на тему «Онлайн система “Терногаз”» (державний реєстраційний номер 0119U102841) (акт про використання результатів дисертаційної роботи від 30 жовтня 2020 р.);

- у навчальному процесі Західноукраїнського національного університету на кафедрі комп'ютерних наук при викладанні дисциплін: «Архітектура та проектування програмного забезпечення», «Веб-програмування», «Засоби програмування баз даних і знань», «Якість програмного забезпечення та тестування» (акт про впровадження в навчальний процес від 06 листопада 2020 р.).

Особистий внесок здобувача. Усі результати, викладені в дисертаційній роботі, отримані автором самостійно. У друкованих працях, опублікованих у співавторстві, автору належать такі результати:

[1] обґрунтовано базові показники оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту;

[2] розроблено метод налаштування та моделювання оцінки ефективності функціонування інформаційних веб-ресурсів;

[3] розроблено метод структурування, розпізнавання і порівняння текстового контенту веб-ресурсів із елементами машинного навчання та із використанням автоматизованого структурування контенту, виявлення недостовірної, неточної чи неактуальної інформації;

[4] удосконалено методи автоматизованого контент-аналізу, пошуку застарілої та некоректної інформації з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, створеними як спосіб виявлення подібного контенту на інших веб-ресурсах;

[5] розроблено архітектури прикладних програмних систем, що виконують функції автоматизованого пошуку застарілої та некоректної інформації за рахунок наявності модулів розпізнавання і порівняння текстового контенту.

Основні положення та результати дисертаційної роботи у наведених працях викладені в повному обсязі.

Апробація результатів дисертації. Основні положення і результати дисертаційної роботи презентовано на 4 конференціях та багатьох наукових семінарах, зокрема:

- Всеукраїнська школа-семінар молодих вчених і студентів «Сучасні комп'ютерні інформаційні технології» ACIT'2016 – Тернопіль – 2016;

- XIII I nternational Conference Perspective Technologies and Methods in MEMS Design - Polyana – 2017;
- IX International Conference on “Advanced Computer Information Technologies” ACIT’2019 – Ceske Budejovice, Czech Republic – 2019;
- 2020 10th International Conference on Advanced Computer Information Technologies (ACIT) – Deggendorf, Germany – 2020.
- Семінари які проводилися на кафедрі комп’ютерних наук ЗУНУ.

Публікації. За результатами дисертаційного дослідження опубліковано 7 наукових праць загальним обсягом 8 сторінок, зокрема 2 статті у фахових наукових виданнях [3] [4] , 1 з яких входить до міжнародних наукометричних баз Web Of Science [4].

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота виконувалася в межах пріоритетного напряму розвитку науки і техніки «Інформаційні та комунікаційні технології», визначеного Законом України «Про пріоритетні напрями розвитку науки і техніки (№2623-III від 11.07.2001 р. в редакції від 16.01.2016 р.), а також згідно з планом науково-дослідних робіт кафедри комп’ютерних наук Тернопільського національного економічного університету протягом 2016 – 2019 років. Основні результати дисертаційного дослідження отримано в межах виконання таких тем:

- госпдоговірної науково-дослідної роботи на тему «Онлайн система “Терногаз”» (державний реєстраційний номер 0119U102841), у якій автором розроблено програмну систему для виявлення некоректної та недостовірної інформації на інформаційних веб-ресурсах

Усі вищезгадані роботи виконувалися за безпосередньої участі автора.

Структура та обсяг роботи. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел із 140 найменувань та додатків. Загальний обсяг роботи складає 141 сторінок машинного тексту, з них - 124 сторінка основного тексту. Робота містить 48 рисунків і 14 таблиць.

РОЗДІЛ 1
АНАЛІЗ МЕТОДІВ, ЗАСОБІВ ТА ТЕХНОЛОГІЙ ОРГАНІЗАЦІЇ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ

Розвиток інформаційних технологій забезпечує розвиток суспільства, переосмислення його пріоритетів, особливо модернізацію існуючих галузей економіки та створення і стійкий розвиток нових. Цифровізація усіх сфер діяльності людини сьогодні є життєвою необхідністю. Одним із засобів забезпечення потреб суспільства є інформаційні веб-ресурси. Їхня сфера застосування від сайтів візитівок (статичних) організацій до інтелектуальних веб-ресурсів для надання різноманітних послуг. Разом з тим, ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту, а також наявність у ньому застарілої і неактуальної інформації. Тому її виявлення та оновлення є актуальним завданням, вирішення якого лежить у сфері розробки засобів автоматизованого структурування та розпізнавання контенту. Особливої актуальності це питання набуває сьогодні, коли активно розвиваються технології веб.3.0 та створено технології веб.4.0, особливостями яких є широке застосування автоматичних систем контент аналізу із використанням засобів штучного інтелекту для структурування та розпізнавання контенту, а також ґрунтовним підвищенням ефективності інформаційних веб-ресурсів.

Саме ці питання розглянуто нами у даному розділі. Обґрунтовано актуальність науково-прикладної задачі розробки математичного та програмного забезпечення для автоматизованого структурування і розпізнавання контенту із використанням засобів машинного навчання, яке у сукупності забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів, а також здійснено постановки задач дисертаційного дослідження.

1.1. Особливості структурування та організації веб-ресурсів

Розглянемо особливості організації інформаційних веб-ресурсів. Структура сайту є найважливішим інструментом із точки зору SEO. Неправильна структура сайту ускладнює його відвідуваність. Тому при розробці архітектури ресурсу необхідно проаналізувати розміщення кожного розділу та підрозділу, щоб зробити його таким, який би задовільнив потреби користувача і реагував на вимоги пошукових роботів [16-27].

Немає чіткого визначення щодо того, якою є правильна структура. Це залежить від типу веб-сайту, семантичного ядра та цільової аудиторії, тому він завжди індивідуальний. Однак існують еталонні типи конструкцій, а також основні правила його розвитку.

Структура сайту - логічна побудова всіх сторінок ресурсу, схема, за якою розподіляється шлях до папок, категорій, підкатегорій (Рис.1.1).

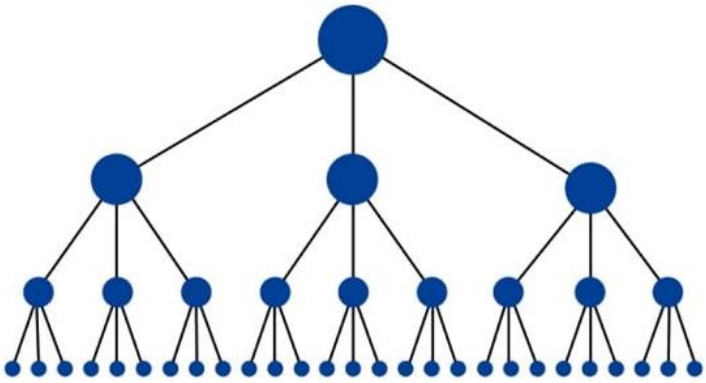


Рис. 1.1. Структура сторінок веб-ресурсу

Пошукові роботи переходять за всіма посиланнями сайту. Якщо структура веб-сайту правильна, пошуковий робот витратить менше часу на аналіз усього ресурсу, що пришвидшить індексацію сайту для ефективного SEO просування. Правильно та логічно розміщені документи на сайті можуть потрапити в індекс пошукової системи вже наступного дня. Це невід'ємна частина для SEO-просування [22-33].

Пошукова система проводить аналогію між сайтом і ставленням користувача до інформаційного ресурсу. Чим привабливіший ресурс для

користувачів, тим привабливіший він і для пошукових систем. Якщо алгоритм Google бачить, що сайт має низький час перебування відвідувача, тобто поганий показник відвідуваності, то сприймає ресурс як неякісний, а це погано вплине на просування та результати пошуку веб-сайту [34-42].

Правильна структура веб-ресурсу фокусує користувача на основних процесах на сайті. Це унеможливлює складність пошуку правильної сторінки (Рис.1.2). В іншому випадку, це незворотно зменшить конверсію, оскільки користувач не знайде того, що шукає, і, відповідно, не стане постійним відвідувачем.

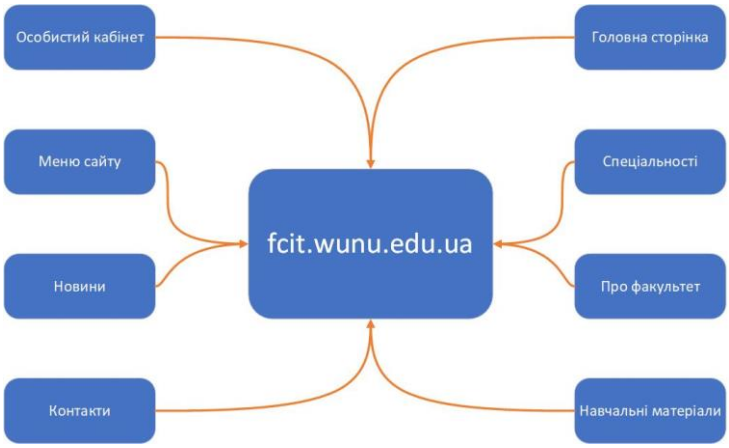


Рис. 1.2. Структура веб-ресурсу http://fcit.wunu.edu.ua/

Якщо вже є семантичне ядро та уявлення про всю інформацію, яка використовуватиметься на веб-ресурсі, формується ієрархічний, логічний і послідовний ланцюжок побудови та подання інформації.

Структура ділиться на рівні, починаючи з першого. Перший рівень - головна сторінка та основні категорії, другий - підкатегорії і так далі. Ієрархічна структура будівництва дозволить користувачеві знайти потрібну інформацію зручніше й швидше.

Шлях від головної сторінки докорінної нам інформації має займати не більше трьох кліків (Рис.1.3).

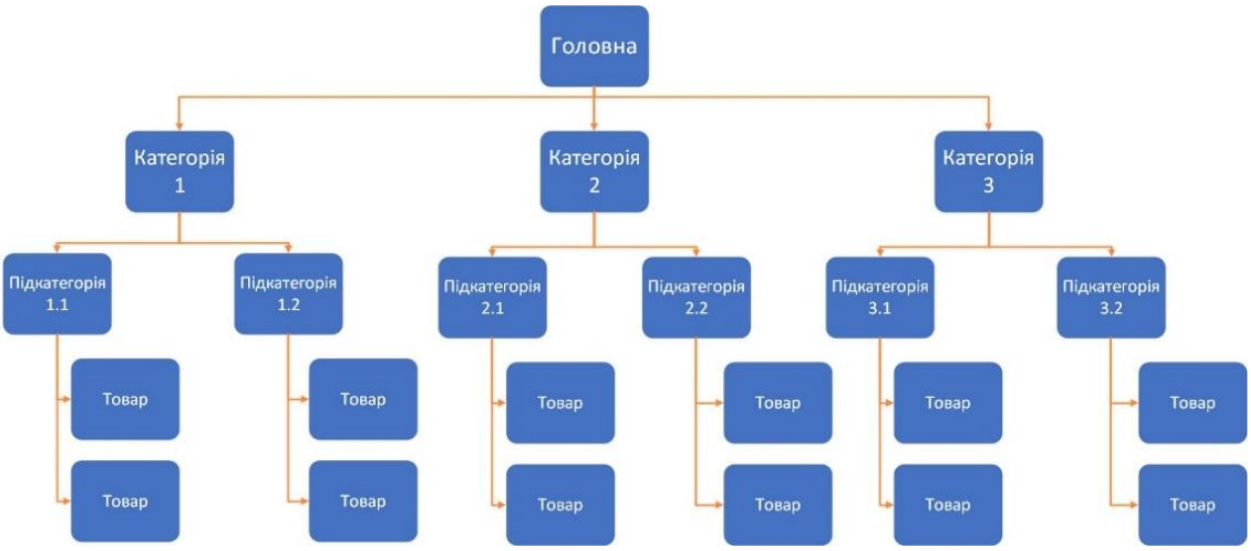


Рис. 1.3. Ієрархічна структура будівництва сторінки докорінної інформації

Однак, такий тип структури сайту не завжди доречний та залежить від його цілей.

Розрізняють наступні типи організація структури веб-сайтів:

1. Алфавітна організація структури - передбачає розміщення інформації в алфавітному порядку. Використовується при структуруванні словників й енциклопедій, є стандартною, але може знайти застосування і в каталогах, метою яких є демонстрація та продаж продукту. Цей тип не завжди зручний як первинний, і, за необхідності, його використовують як додаткову інформацію для побудови інформації в алфавітному порядку, тим самим спрощуючи пошук.

2. Хронологічна організація - запит споживача пов’язаний з конкретною датою. Найчастіше цей метод використовується при розміщенні новин, прес-релізу або інформації на сайті та дозволяє швидко орієнтуватися у часі, вибираючи інформацію, яка його цікавить. Обмеження для споживача - розібратися в чітких часових рамках: даті, місяці, періоді.

3. Географічна організація - використовується якщо потрібно сортувати за місцем для доступу до інформації. Ця класифікація не завжди зручна, оскільки існує досить велика аудиторія клієнтів і читачів, які можуть не

знати географії певної області, тому для них буде важко відшукати потрібну інформацію.

4. Тематична організація - використовується інтернет-магазинами, інформаційними порталами з великою кількістю інформації та багатьма іншими сайтами. Така організація дозволяє швидко пересуватися веб-сайтом, вибираючи в один клік лінк, який цікавить користувача. Тематична організація оптимальна для SEO, оскільки веде від загальних понять до приватних концепцій і дозволяє реалізувати все семантичне ядро на сайті.

5. Організація яка націлена на цільову аудиторію - дозволяє одночасно охоплювати всі ідентифіковані групи цільової аудиторії з різними потребами. Забезпечує зручність для користувача, а також дозволяє створити правильні умови для використання ресурсу. Зокрема, це можуть бути додаткові функції веб-сайту, необхідні для певної цільової групи, іншого дизайну, навіть іншого меню для різних цільових аудиторій на основі їхніх інтересів тощо.

Ключовими вимогами до структури сайту є структура інформації на ресурсі, яка одночасно впливає на два показники: коефіцієнт користувача й ефективність, швидкість ранжування ресурсу пошуковою системою.

Тому розробка схеми повинна одночасно відповідати вимогам як споживача, так і пошукової системи.

Види структур інформаційних веб-ресурсів:

1. Структура сайту візитівки - простий ресурс із декількох сторінок для презентації споживачеві інформації про компанію. Структура цього сайту часто найпростіша - лінійна, з мінімальним вкладенням, а необхідний мінімум сторінок на сайті визначається виходячи з теми (Рис.1.4).

Рис. 1.4. Структура сайту візитівки

2. Структура комерційного веб-сайту - проект, який найчастіше спрямований ну тільки на продаж, але й на детальну презентацію замовнику самої компанії. Відповідно, структура повинна містити якомога більше інформації (Рис.1.5).

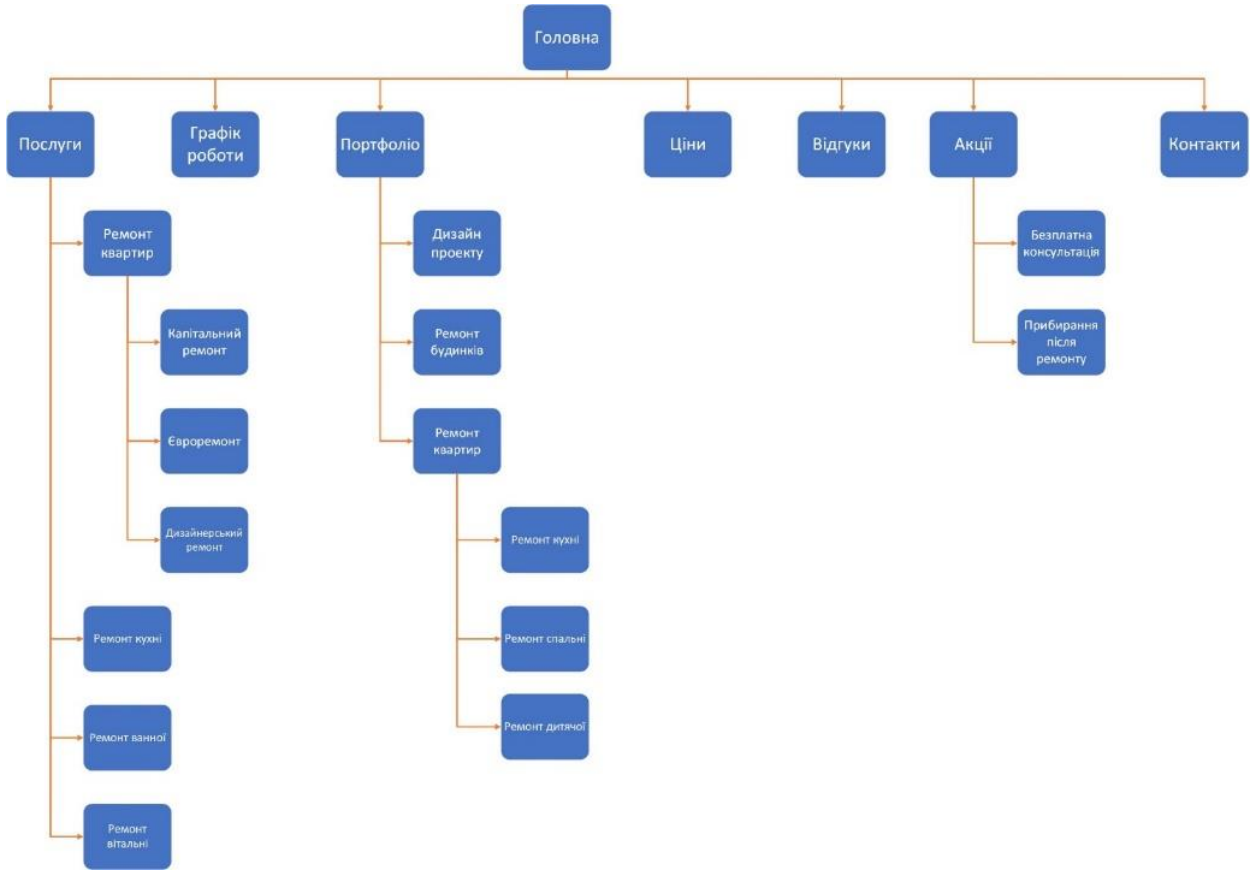


Рис. 1.5. Структура комерційного веб-сайту

У цьому випадку структура здається складною й заплутаною, хоча насправді, щоб перейти від категорії до підкатегорії 3 рівня, потрібно зробити всього 2 кліки. Такий підхід нагадує структуру однострічкового сайту.

3. Структура інформаційного порталу - міські інформаційні портали поєднують новини у розділах спорту, подій, інцидентів тощо (Рис.1.6).

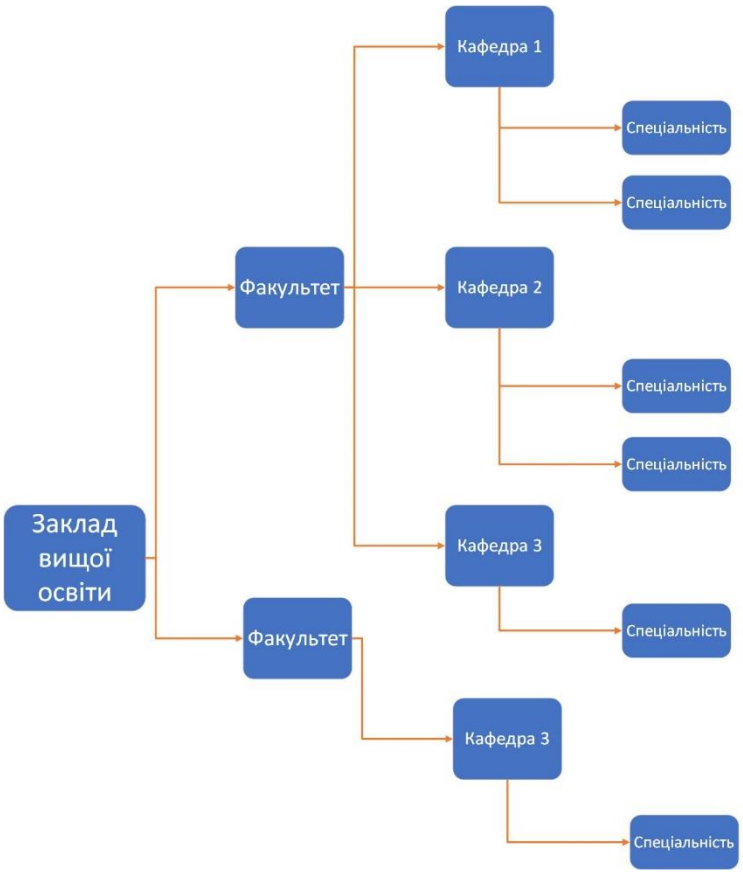


Рис. 1.6. Структура інформаційного порталу

Правильна структура сайту - запорука його успіху й ефективності. Прямий вплив на зручність пошуку у поєднанні із семантичним ядром сприяє досягненню хороших результатів із точки зору SEO. Будь-які, навіть невеликі, помилки, зроблені при розробці схеми, можуть створити критичні проблеми для сайту.

Тепер розглянемо особливості представлення та оцінки веб-контенту.

Існує декілька ключових показників, які визначають загальну релевантність вмісту:

1. Для визначення актуальності чи недостовірності контенту пошукові системи звертають увагу на дату створення сторінок. На перших позиціях розміщені ті сайти, інформація яких була опублікована трохи пізніше.

2. Важливим є оновлення контенту на сайті, що також впливає на пошукові системи при визначенні його актуальності. Щоб постійно мати важливий і корисний контент на ресурсі, недостатньо просто змінити кілька слів у статтях або навіть видалити одне речення, оскільки численні конкуренти, які

змінюють зміст на пару абзаців, обов'язково обійдуть ресурс у процесі пошуку.

3. Контент потрібно оновлювати часто. Частота вимірюється в днях, а не місяцях, упродовж багатьох років.

4. Пошукові системи приділяють увагу свіжому контенту, численним новостворених сторінкам. Зважаючи на це, необхідно забезпечити збільшення загальної кількості сторінок щороку приблизно на 30%.

5. Загальна поведінка користувача відіграє певну роль, яка також показує, наскільки релевантними є вміст. Якщо оновлення контенту ресурсу викликає серйозне пожвавлення користувачів до нього, то вони набагато довше залишаються на сторінках, відповідно, можна судити про актуальність інформації.

6. Якщо ресурс згадується все частіше, якщо він з’являється у все більшій загальній кількості посилань на ресурс із інших сайтів, репостів, лайків у соціальних мережах, пошукові системи роблять висновок, що сайти містять надзвичайно актуальний контент.

Актуальність та новинність описуються такими методами для підвищення цих параметрів:

1. Зміна документу, домену або веб-сторінки. Сайти, які додають нові та створюють абсолютно новий контент, привертатимуть увагу цільової аудиторії.

2. Внесення змін до важливих областей документів, які показують важливість сторінки. Менш важливим вмістом є реклама та навігація. Найважливіша інформація, як правило, знаходиться на головній сторінці ресурсу.

3. Збільшення кількості посилань та частоти їх додавання. На новизну сайтів впливають посилання із найсвіжіших сторінок ресурсу. Новинними є також ті ресурси, на які всі постійно посилаються. Якщо сайт кардинально зміниться сам і своє розташування, новий якір також буде змінений. Пошукова система визначає, що ресурс настільки змінився, що старий якірний текст більше не актуальний. Це повністю знецінює всі старі зв’язки [43-45].

Таким чином, ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити: продуктивність

апаратного і програмного забезпечення, використаного джерелами, способи організації джерела, зокрема структурованість контенту та складність запиту.

Разом з тим ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту, а також наявність у ньому застарілої та неактуальної інформації.

1.2. Огляд відомих систем контент аналізу

1.2.1. Веб-технології як основа побудови систем аналізу контенту з елементами машинного навчання

Одним із відомих методів аналізу текстової інформації є контент-аналіз – стандартна методика дослідження, предметом якої є аналіз змісту текстових масивів і продуктів комунікативної кореспонденції (наприклад, коментарі, форуми, електронне листування, статті тощо) [42-50].

Поняття контент-аналізу не має однозначного визначення, що породжує проблему: системи, побудовані на основі різних підходів до контент-аналізу, в загальному випадку несумісні [50-52].

Таблиця 1.1.

Аналіз методів контент-аналізу

Назва методу	Показники		
	Складність реалізації	Часові параметри	Можливість програмної інтеграції
Методи об’єктно - якісного та систематичного дослідження змісту засобів комунікації (Д. Джері, Дж. Джері)	Висока	Великі часові витрати на формалізацію	Часткова інтеграція в існуюче ПЗ
Систематичне кількісне опрацювання, оцінювання та інтерпретація форми і змісту інформаційного джерела (Д. Мангейм, Р. Рич)	Середня	Великі часові витрати на формалізацію	Інтеграція на рівні окремого ПЗ

Продовження таблиці 1.1.

Якісно-кількісний метод дослідження документів (характеризується об'єктивністю висновків і строгістю процедури) та квантифікаційного опрацювання тексту з подальшою інтерпретацією результатів; предмет дослідження – проблеми соціальної дійсності, які висловлюються і приховуються у документах, та внутрішні закономірності самого об`єкту дослідження (В. Іванов)	Висока	Великі часові витрати на формалізацію	Інтеграція на рівні окремого ПЗ
Пошук у тексті визначених змістовних понять (одиниць аналізу), виявлення частоти їх появи та співвідношення із змістом всього документа (Б. Краснов)	Середня	Середні часові витрати на формалізацію	Часткова інтеграція в існуюче ПЗ

Контент-аналіз - метод якісно-кількісного аналізу змісту документу, за допомогою якого можливе більш глибоке осягнення текстової реальності. Основними складовими дослідження контент-аналізу [59-62]:

1. Виокремлення контенту.
2. Структурування - гіпотеза структурування, сегментація, поділ текстів (повторення в усіх/ряді текстів) у всьому досліджуваному масиві .
3. Формалізація – використання однотипних сегментів та застосування високого ступеня формалізації, визначених правил та формальних алгоритмів у аналітичних процедурах.
4. Реферування – формальний поділ текстів та визначення окремих елементів для наступного пошуку із застосуванням аналітико-синтетичної процедури.
5. Аналіз – застосування методів математичної статистики та методів теорії ймовірності для аналізу тексту.

Контент-аналіз – це кількісно-якісний аналіз текстової інформації та текстових масивів з метою подальшої змістовної інтерпретації отриманих

кількісно-якісних закономірностей. Важливим питанням контент-аналізу є його автоматизація у спосіб використання спеціальних програмних систем [63-67].

У сучасних веб-технологіях реалізовано деякі процедури, які дають можливість інтегрувати системи контент-аналізу.

Веб - це найбільша конструкція, що перетворюється, її ідея представлена Тімом Бернерсом-Лі в 1989 р. За останні два десятиліття в Інтернеті та суміжних технологіях було досягнуто значного прогресу. Веб 1.0 – як мережа пізнання, веб 2.0 – як мережа спілкування, веб 3.0 – як мережа співпраці та веб 4.0 – як мережа інтеграції. Вони представлені, наприклад, як чотири покоління мережі з моменту появи Мережі [72-74].

Веб 3.0, або семантична мережа зменшує завдання та рішення, які виконує людина. Загалом, веб-версія 3.0 включає дві основні платформи: семантичні технології та соціально обчислювальне середовище. Семантичні технології представляють відкриті стандарти, які можна застосовувати в Інтернеті. Соціальне обчислювальне середовище забезпечує співпрацю людини-машини та організовує велику кількість соціальних веб-спільнот [117-119].

Веб 4.0 – паралельне читання-запис-виконання мереж із інтелектуальною взаємодією, але точного визначення її досі немає. Інтернет 4.0 також відомий, як симбіотична павутина, в якій людський розум і машини можуть взаємодіяти в симбіозі [75-78].

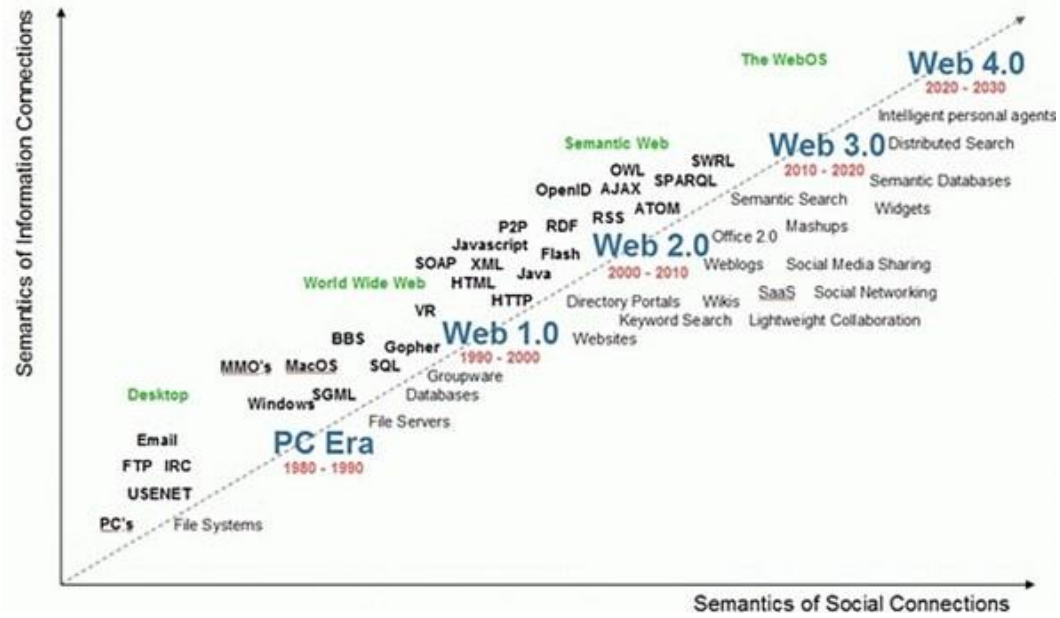


Рис. 1.10. Порівняння концепцій Веб 1.0, Веб 2.0, Веб 3.0 та Веб 4.0

Джон Маркофф із New York Times запропонував веб 3.0 в якості третього покоління Інтернету в 2006 році. Основна ідея веб 3.0 полягає у визначенні структурних даних та їх зв'язків для ефективнішого виявлення, автоматизації, інтеграції та повторного використання у різних додатках. Веб 3.0 здатний поліпшити управління даними, підтримати доступність мобільного Інтернету, імітувати креативність та інновації, допомагати організовувати співпрацю у соціальних мережах [100-102].

Веб 3.0 також відомий як семантична мережа. Семантичну павутину придумав Тім Бернерс-Лі, винахідник Всесвітньої павутини. У консорціумі Всесвітньої павутини (W3C) існує спеціальна команда, яка працює над удосконаленням, розширенням та стандартизацією системи, мови, публікації. Семантична мережа - це мережа, яка може продемонструвати речі в тому підході, який може зрозуміти комп'ютер. Головною важливою метою семантичної мережі є зробити мережу читабельною для машин, а не тільки для людей. В цьому сенсі структурування та розпізнавання контенту є важливим завданням в межах зазначених технологій [104-106].

Разом з тим, формальний опис контенту, залишається актуальним, не залежно від стрімкого росту розглянутих технологій, а машинне навчання для його розпізнавання є ключовим процесом у виявлені недостовірної інформації, чи її неактуальності.

Алгоритми машинного навчання можна класифікувати як навчання під наглядом або без нагляду. При контрольованому навчанні навчальні приклади складаються із шаблонних пар вводу-виводу. Метою алгоритму навчання є прогнозування вихідних значень нових прикладів на основі їх вхідних значень. У навчанні без нагляду, приклади містять лише схеми введення, і жоден явний цільовий результат не пов'язаний з кожним входом. Алгоритм навчання повинен узагальнюватись за вхідними шаблонами, щоб виявити вихідні значення [4, 31-33].

Таблиця 1.3.1

Класифікація методів та програм

	Характеристики	Будь-які дані	Текстові дані	Дані, пов’язані з Інтернетом
Призначення	Отримання із відомих даних або документів	Пошук даних	Пошук інформації	Пошук в інтернеті
	Пошук нових зразків знань, раніше невідомих	Отримання даних	Отримання тексту	Отримання даних з інтернету

За минулі десятиліття було розроблено багато систем машинного навчання. Виділено три класи технік машинного навчання [50-52]:

1. Навчання на основі прецедентів;
2. Нейронні мережі;
3. Еволюційні алгоритми.

Опираючись на ці дві класифікації та огляд галузі, прийняли подібні рамки та визначилися наступні п’ять основних парадигми [95]:

1. Ймовірнісні моделі;
2. Навчання на основі прецедентів та індукція правил;
3. Нейронні мережі;
4. Еволюційні моделі;
5. Аналітичне навчання та нечітка логіка.

Використання ймовірнісних моделей було однією з найперших спроб виконувати машинне навчання, найпопулярнішим прикладом якого є Байєсівський метод. Початок досліджень – розпізнавання зразків (Duda &Hart, 1973), цей метод часто застосовувався для класифікації різних об'єктів, де заздалегідь визначені класи на основі набору функцій [137-138].

Навчання на основі прецендентів – класифіковане відповідно до основної стратегії навчання, наприклад, навчання навчити, навчання за інструкцією, навчання за допомогою аналогії, навчання на прикладах та навчання на відкриттях (Carbonell, Міхальські та Мітчелл, 1983; Коен і Фейгенбаум, 1982). Серед них, навчання на прикладах, здається, найбільш перспективною –

навчання на основі прецендентів для виявлення знань та аналізу даних. Реалізується шляхом застосування алгоритму, який намагається викликати загальний опис концепції, який найкраще описує різні класи навчальних прикладів. Розроблено численні алгоритми, кожен із яких використовує один або більше методів створення ідентифікаційних шаблонів, які корисні в створенні опису концепції. Алгоритм побудови дерева рішень ID3 Квінлана (Quinlan, 1983) та такі варіації, як C4.5 (Quinlan, 1993), мають стати одними з найширше використовуваних у навчанні на основі прецендентів [138-140].

Враховуючи набір об’єктів, ID3 створює дерево рішень, яке намагається правильно класифікувати всі об’єкти. На кожному кроці алгоритм знаходить атрибут, який найкраще поділяє об'єкти на різні класи, мінімізуючи ентропію (невизначеність інформації). Адже, всі об'єкти були класифіковані, або використовують усі атрибути, результати представлені деревом рішень чи набором правил.

Штучні нейронні мережі намагаються досягти людських характеристик шляхом моделювання нервової системи людини. Хоча знання представлені символьними описами, такі як дерево рішень та правила при навчанні на основі прецендентів, знання засвоюються та запам'ятовуються мережею взаємопов'язаних нейронів, зважених синапсів та порогових логічних одиниць (Lippmann, 1987; Rumelhart, Hinton, & McClelland, 1986). На основі навчальних прикладів можна використовувати алгоритми навчання для регулювання ваги з'єднання в мережі, щоб вона могла передбачити або класифікувати невідомі приклади правильно. Тоді алгоритми активації над вузлами використовується для отримання концепцій та знань із мережі (Belew, 1989; Chen & Ng, 1995; Квок, 1989).

Разом із тим, нейронні мережі вимагають формалізації представлення контенту та його кодування для використання в задачах автоматизованого аналізу контенту [120-122].

Інший клас алгоритмів машинного навчання складається з еволюційних алгоритмів, які опираються на аналогії із природніми процесами Фогель (1994) виділяє три категорії еволюційних алгоритмів: генетичні алгоритми, еволюційні

стратегії та еволюційне програмування. Генетичні алгоритми виявилися популярними та успішно застосовані до різних задач оптимізації. Вони базуються на генетичних принципах (Goldberg, 1989; Michalewicz, 1992).

Аналітичне навчання представляє знання, як логічні правила і виконує міркування щодо цих правил для пошуку доказів. Докази можна скласти більш складні для вирішення проблем із невеликою кількістю пошукових запитів. Традиційні аналітичні системи навчання залежать від жорстких обчислювальних правил, як правило, не існує чіткого розмежування між цінностями та класами в реальному світі. Нечіткі системи дозволяють значенням "false" або "true" працювати в діапазоні дійсних чисел від нуля до одиниці (Zedah, 1965). Нечіткість вміщує неточність та приблизні міркування [79-84].

Разом із тим, генетичні алгоритми, подібно як і нейронні мережі вимагають формалізації представлення контенту та його кодування для використання в задачах автоматизованого аналізу контенту.

Виходячи із викладеного вище, серед машинних методів навчання для структурування та виявлення некоректної, застарілої чи неактуальної інформації найбільш легко адаптувати метод навчання на основі прецедентів.

1.2.2. Огляд відомих спеціалізованих систем контент-аналізу

Проведемо огляд відомих систем контент-аналізу. Системи контент-аналізу поділяються за типами.

Перший тип – повністю автоматизовані пакети для контент-аналізу. Включають в себе розроблені програм аналітичні словники, в які, зазвичай, неможливо або досить важко внести зміни. Такі програми застосовуються для перевірок лінгвістичних гіпотез, закладених в основу програмних словників та аналітичних схем. Адже сучасний рівень розвитку "штучного інтелекту" не дозволяє аналітику передавати програмі повний контроль над процесом інтерпретації масиву документів. Жодна сучасна програма не може розуміти текст в людському значенні цього поняття [85-88].

Другий тип програм – надзвичайно прості та універсальні пакети підрахунку частот різних слів в текстах. До недоліків таких програм слід

віднести відсутність категоризації підрахованих слів, що призводить до викривлення результатів в бік окремо беззмістовних слів, що належать до службових частин мови. Крім того, різні форми змістовних слів рахуються як різні слова, що не дозволяє поставити знак рівності між одиницями підрахунку та одиницями значень. Останній із згаданих недоліків був частково розв'язаний розробниками через процедуру лематизації (виділення незмінних частин слів). Однак подібних програм з лематичним вдосконаленням для слов'янських мов, які мають надзвичайно гнучку морфологію, яка передбачає досить відмінні форми для різних родів, чисел, відмінків і т.д., не існує [89-94].

Третій тип програм – напівавтоматичні пакети для ручного кодування текстових даних. Вони спрощують деякі технічні аспекти кодування: користування ними аналогічне альтернативі "писати олівцем чи писати на комп'ютері". Дані комплекси позбавляють кодувальника постійно звертатись до кодувальної таблиці, упорядковують результати кодування багатьох кодувальників, зберігають дані в зручній для пошуку і маркування формі, візуалізують знайдені зв'язки. Дозволяють заощадити певний час та підвищують ефективність, однак у випадку масивів інформації, що вимірюється сотнями і тисячами сторінок, дане програмне забезпечення відчутно не допоможе. Тому такі програми найчастіше використовуються для упорядкування якісного контент-аналізу та для контент-аналітичних кейс-стаді [96-99].

Далі розглянемо стислі характеристики програмних систем контент-аналізу.

ЛЕКТА – створює багатовимірний контент-аналіз текстових масивів. На початковому етапі допомагає скласти словник контент-аналізу як на основі частотності, так і на основі заздалегідь створеній системі категорій. Дозволяє розбити тексти на рівні за обсягом фрагменти. Дозволяє об'єднати одиниці рахунку і фрагменти текстів в групи з використанням факторного аналізу. Таким чином, дослідник отримує в своє розпорядження чітку структуру характеристик досліджуваного інформаційного простору, обґрунтовану принципом частотності включених в словник лексем. Після цього слід якісна інтерпретація отриманих тематичних блоків [107-108].

JFreq – створює матриці частот вживання слів у масиві, використовується для проведення контент-аналізу і працює з більшістю мов світу. Чи не працює з японським, китайським і тайським мовами через принципово відмінною від більшості мов лінгвістичної системи цих мов. Програма дозволяє виключити з масиву нечитабельним символи і знаки не входять в алфавітну базу даних [109-111].

Concordance – програма, яка використовується для проведення контент-аналізу електронних документів, В ній можна створювати списки пов'язаних одиниць рахунку, індексів, слів, при роботі над електронним текстом. Дозволяє обробляти великі масиви. Дає можливість переглядати кореляції між словами, що входять в словник контент-аналізу. Результати роботи легко розмістити в Інтернет за допомогою вбудованих інструментів програми [112-114].

WordStat – модуль аналізу тексту, призначений спеціально для обробки матеріалів, таких як журнальні статті, літературні твори, інтерв'ю. Як і інші аналогічні програми дозволяє створювати категоріальний апарат і словник контент-аналізу. Подальший аналіз може бути проведений за допомогою створення і розрахунку перехресних таблиць, а також KWIS-методу. Пакет дозволяє працювати і з більш складними методами статистичного аналізу, такими як кластеризація і багатовимірне шкалювання. Створені категоріальні апарати і словники схеми застосовуються до інших текстовим масивів [115].

SALT – програмне забезпечення, що аналізує вміст текстового масиву. Підтримує роботу з усіма мовами. Визначає середню довжину речення, кількість шуканих слів, загальна кількість слів. Може створювати алфавітний список слів, кодувати текстовий масив відповідно до визначених дослідником кодами. Працює тільки з операційною системою Windows.

QDA Miner є засобом якісного аналізу текстових даних, анотування, отримання та перегляду кодованих даних. Програма дозволяє працювати великим числом документів, що містять як текст, так і числові дані. QDA Miner має широкий спектр пошукових засобів для виявлення кореляцій кодованих даних [116].

У вище розглянутих системах відсутні:

1. Методи моделювання базового показника оцінки ефективності функціонування веб-ресурсів залежно від структурованості контенту, що забезпечують вирішення цілого ряду практичних завдань, що виникають при розробці і впровадженні веб-систем: налаштування параметрів продуктивності, аналіз впливу структурованості контенту, оцінка ефективності функціонування під навантаженням.

2. Формальний опис контенту, що не дає можливості удосконалювати процедури структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання.

1.3. Обґрунтування показника ефективності функціонування інформаційних веб ресурсів

Існуючі сайти поділяються на дві категорії – сайти для електронної комерції та корпоративні. Для оцінки ефективності функціонування веб-ресурсів розглянуто систему показників (метрик). Для сайтів, що належать до категорії електронної комерції, використовуються чотири метрики: конверсія, час перебування на сайті, кількість відмов і кількість переглянутих сторінок [53-58].

Для категорії корпоративних сайтів метрика «конверсія» не може бути використана, однак всі інші показники застосовуються. Показник «час перебування користувача на сайті» є найважливішим показником його ефективності, незалежно від того, до якої категорії відноситься сайт.

Оцінку часу перебування користувача на сайті можна проводити за допомогою спеціальних інструментів. Часто виникає задача оцінки тимчасових характеристик сайту до його створення, тобто в процесі проектування. Зазвичай, проектований сайт орієнтується на високий рівень активності користувача. Оскільки процеси, що відбуваються на сайті, носять імовірнісний характер, для оцінки його тимчасових характеристик імовірним є використання методів аналітичного моделювання і теорії масового обслуговування [123-140].

Сучасні аналітичні моделі масового обслуговування дозволяють обчислювати середній та максимальний час перебування заявок у системі або

мережі. Під максимальним часом перебування заявки у системі або мережі масового обслуговування розуміється час, перевищення якого можливе тільки для деякої, наперед заданої частки заявок. Наприклад, для багатолінійних систем масового обслуговування максимальний час перебування у системі обчислюється за допомогою рішення трансцендентних рівнянь [68-69].

Теорія масового обслуговування дозволяє будувати і досліджувати досить прості моделі функціонування різних систем.

Кожну із сторінок проектованого сайту можна представити в аналітичній моделі у вигляді системи масового обслуговування, що моделює певну затримку (час перебування користувача на сторінці).

Використовуються такі показники:

1. Кількість сторінок веб ресурсу.
2. Інтенсивність вхідного потоку заявок на сторінку ззовні.
3. Частина відвідувачів сайту, для якої обчислюється максимальний час перебування на веб-ресурсі.
4. Середні час перебування користувача на веб-ресурсі.

Оцінка ефективності інформаційних веб-сайтів - ознайомлення з веб-документацією пошукової системи. Основним завданням інтернет-ресурсу є надання користувачеві потрібної інформації [70-71].

Виділено такі критерії оцінки якості веб-ресурсів:

- зовнішній вигляд веб-документу, дизайн веб-сайту є головним і найчастіше єдиним критерієм його якості;

- структура та навігація - керування вмістом веб-сайтів, а також різні сервіси. Наприклад, «Кошик замовлень», «Пошук», «Форма реєстрації користувача». Функціональність веб-сайту - простота та зручність, оскільки працюватимуть із ним користувачі, які можуть нічого не знати про веб-дизайн та HTML;

- основним критерієм оцінки якості є зручність використання веб-сторінок. За допомогою дружнього стосовно користувачів сайту можна завоювати довіру, а поганий сайт тільки їх відштовхне.

Сайти, створені відповідно до правил оцінки якості та зручності

використання веб-сторінок, найпопулярніші та мають більше високу статистику відвідувань [8-9].

Таким чином, навігація веб-сайту повинна бути простою та зручною, веб-сторінки мають завантажуватися швидко.

Розглянуто системи для оцінки ефективності ресурсів.

PageSpeed Insights - це інструмент, що вимірює швидкість завантаження сторінки, аналізує параметри і дає рекомендації щодо покращення. Результат перевірки - оцінка за стобальною шкалою. Багато оптимізаторів і експертів не радять повністю орієнтуватися на цю оцінку, оскільки немає точних даних, що вона безпосередньо впливає на ранжування сайту.

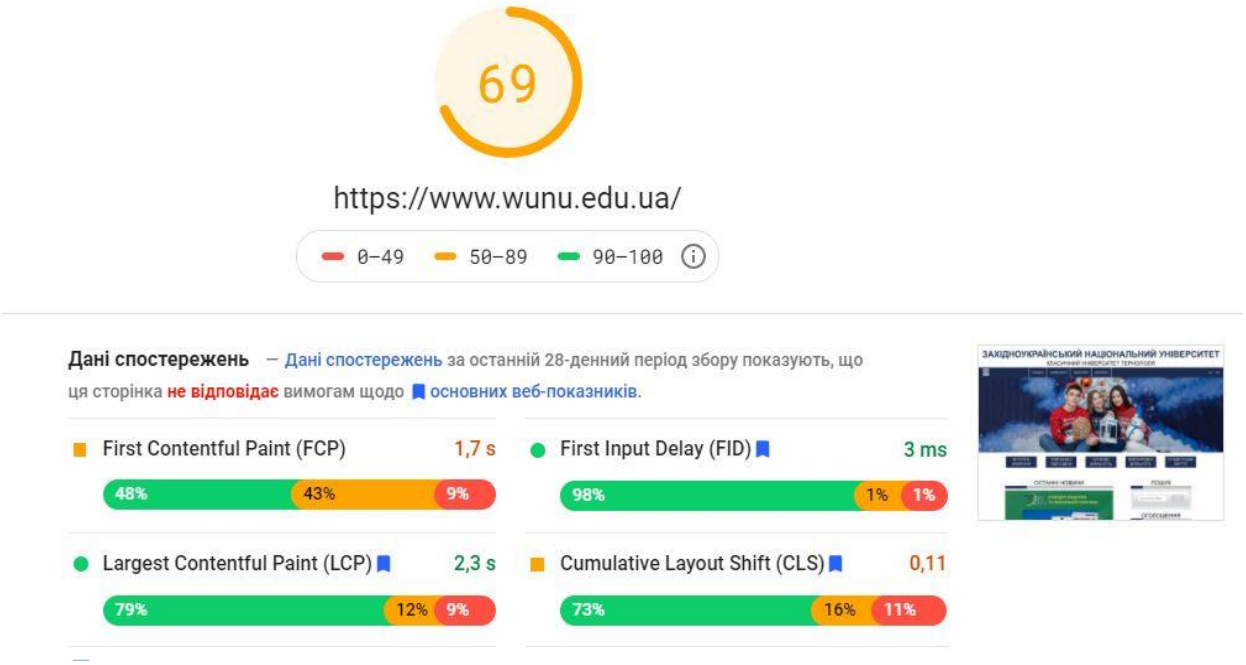


Рис. 1.1. Сторінка системи PageSpeed Insights

Інструмент показує результати для мобільних і десктопних пристроїв й аналізує сторінку за такими параметрами:

1. Перша візуалізація вмісту.
2. Затримка в завантаженні даних.
3. Індекс швидкості.
4. Період до повного завантаження.
5. Перший простій центрального процесора.

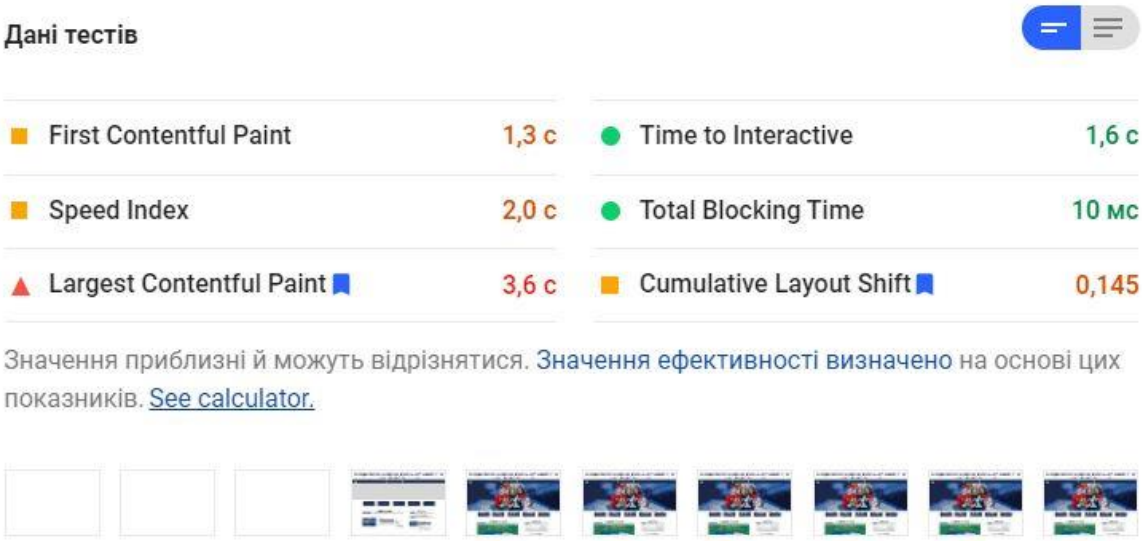


Рис. 1.2. Сторінка системи PageSpeed Insights визначання ефективності веб-ресурсу

Цей інструмент користується великою популярністю у плані зменшення розмірів скриптів, зображень та інших елементів, що корисно для підвищення швидкості завантаження сайту. Використання кешу браузера, рекомендований сервісом у деяких випадках, теж дуже важлива частина прискорення ресурсу [10-11].

Tools.pingdom.com – інструмент для оцінки швидкості роботи веб-ресурсів. Відображає кожний запит, що формується при зверненні користувача до сайту. Інколи на сайтах є безліч зайвих запитів, яких візуально не видно, а фізично вони присутні.

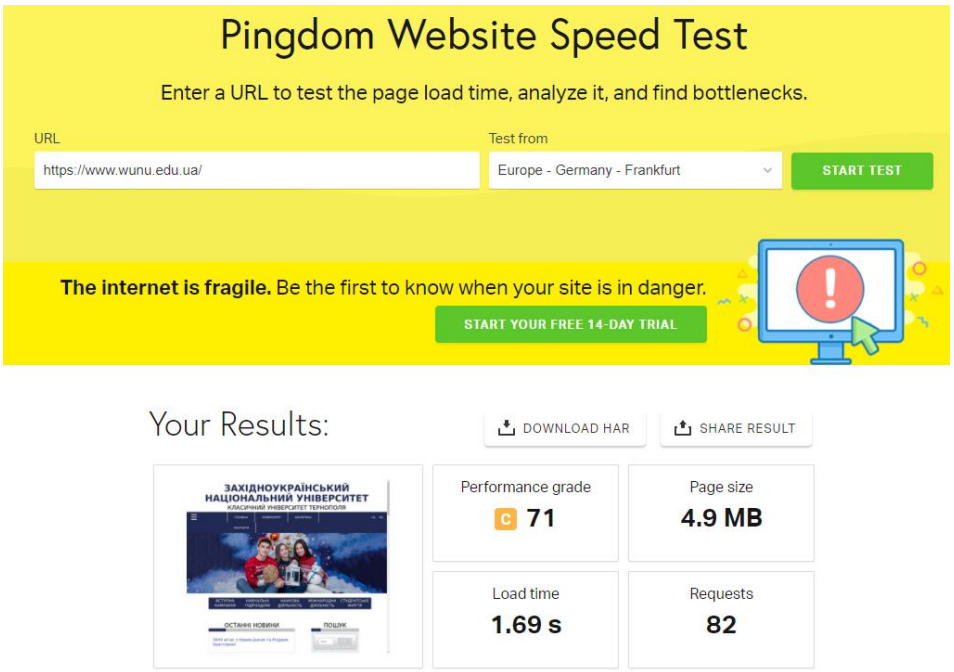


Рис. 1.3. Сторінка системи Tools.pingdom.com

1. Вибираємо сервер, звідки тестуємо сайт. Це необхідно, бо якщо сервер розташований десь в Америці, то дані будуть не точні.

2. Загальне значення показника веб-сайту. Нижче представлено детальний опис.

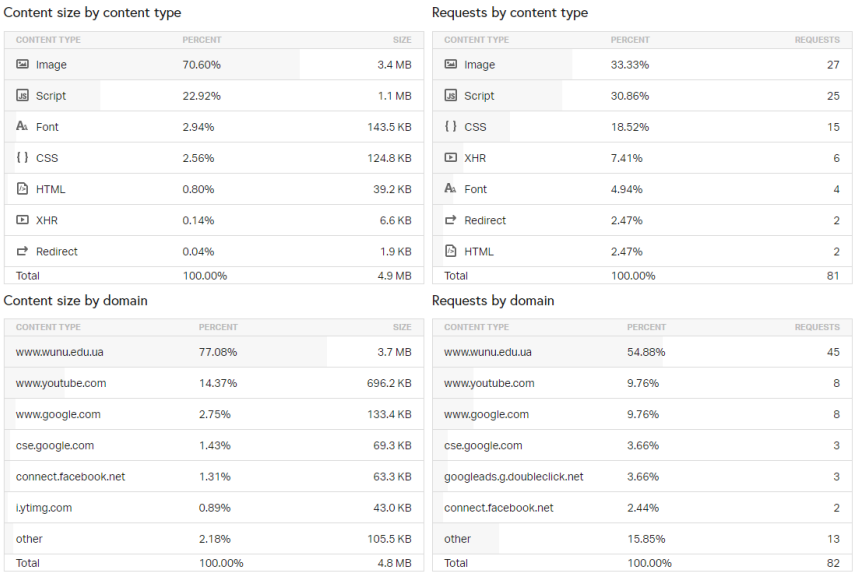
Рис. 1.4. Сторінка системи Tools.pingdom.com із загальними значення показника веб-сайту

3. Швидкість завантаження веб-ресурсу. Чим швидше працює ресурс, тим менше від нього відмовиться людей. Оптимальна швидкість завантаження – до 2 секунд. Ідеальна швидкість прогрузки сайту – менше 1 секунди. Якщо сайт вантажиться довше 3 секунд, є привід задуматися над оптимізацією сайту.

4. Розмір сторінки, яка тестувалася 4.9 MB, що означає непогано. Це дуже важлива метрика. Веб-сайт швидко працює, але через велику кількості картинок на ньому, якщо розмір сторінки містить більше 10 MB інформації, якщо у користувача обмежена швидкість та обмежений об'єм трафіку, тоді є можливість втратити відвідуваність.

Наприклад, підключивши LazyLoad для оптимізації завантаження видимого змісту, не завантажуватиметься інформація, що знаходиться за межами екрану. Карусель товарів, в якій 100 зображень, але відображається перші 4. Решта 96 – приховані, але унаслідок того, що вони технічно завантажуються, уповільнюється відображення сторінки сайту [12-13].

5. Кількість звернень до сервера. Нижче представлено розгорнуту інформацію про те, які конкретно зображення з'являються:



Content size by content type

CONTENT TYPE	PERCENT	SIZE
Image	70.60%	3.4 MB
Script	22.92%	1.1 MB
Font	2.94%	143.5 KB
CSS	2.56%	124.8 KB
HTML	0.80%	39.2 KB
XHR	0.14%	6.6 KB
Redirect	0.04%	1.9 KB
Total	100.00%	4.9 MB

Requests by content type

CONTENT TYPE	PERCENT	REQUESTS
Image	33.33%	27
Script	30.86%	25
CSS	18.52%	15
XHR	7.41%	6
Font	4.94%	4
Redirect	2.47%	2
HTML	2.47%	2
Total	100.00%	81

Content size by domain

CONTENT TYPE	PERCENT	SIZE
www.wunu.edu.ua	77.08%	3.7 MB
www.youtube.com	14.37%	696.2 KB
www.google.com	2.75%	133.4 KB
cse.google.com	1.43%	69.3 KB
connect.facebook.net	1.31%	63.3 KB
i.ytimg.com	0.89%	43.0 KB
other	2.18%	105.5 KB
Total	100.00%	4.8 MB

Requests by domain

CONTENT TYPE	PERCENT	REQUESTS
www.wunu.edu.ua	54.88%	45
www.youtube.com	9.76%	8
www.google.com	9.76%	8
cse.google.com	3.66%	3
googleads.g.doubleclick.net	3.66%	3
connect.facebook.net	2.44%	2
other	15.85%	13
Total	100.00%	82

Рис. 1.5. Сторінка системи Tools.pingdom.com із представленою кількістю звернень до сервера

На вищенаведеному рисунку представлено такі показники: 25 запитів – це JS і 15 запитів - CSS стилі, виходить із 81 запиту. Також представлено самі запити, шлях до них і час завантаження кожного із них.

GTmetrix – це безкоштовна програма, що перевіряє швидкість веб-сторінки та подає детальний звіт. Вона генерує оцінки продуктивності веб-

сторінок та надає рекомендації щодо їх покращення.

Особливості:

1. Тестувати веб-сайтів у різних країнах, а також у браузерах.
2. Надає короткий зміст ключових показників ефективності.
3. Можливість відстежувати ефективність веб-сайту за допомогою моніторингу та бачити його за допомогою графіків.
4. Здатність перевірити веб-сторінку на різних модельованих пристроях.
5. Дозволяє відтворювати завантаження веб-сторінок із відео.

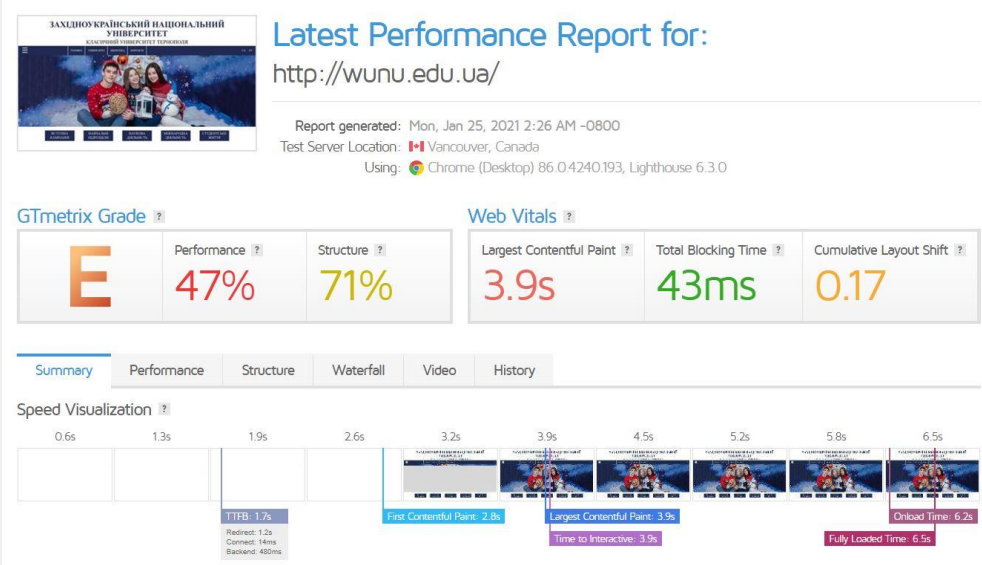


Рис. 1.6. Сторінка системи GTmetrix

Performance Metrics – показник ефективності Lighthouse, зафіксований GTmetrix, з нашими спеціальними аудитами, параметрами аналізу, специфікаціями браузера та обладнання.

Рис. 1.7. Сторінка системи GTmetrix - показник ефективності

Structure – відображає, наскільки добре побудована ваша сторінка для оптимальної роботи.

IMPACT	AUDIT	
Med-High	Avoid enormous network payloads	Total size was 4,583 KiB
Med	Serve static assets with an efficient cache policy	48 resources found
Med	Avoid chaining critical requests	20 chains found
Med-Low	Use a Content Delivery Network (CDN)	43 resources found
Med-Low	Avoid multiple page redirects	Potential savings of 1,180 ms

Рис. 1.8. Сторінка системи GTmetrix структурованість сторінки

Web Vitals – невеликий набір основних показників, які вказують на те, чи швидко надається доступ відвідувачам і (як називає Google) чудовий досвід [14-15].

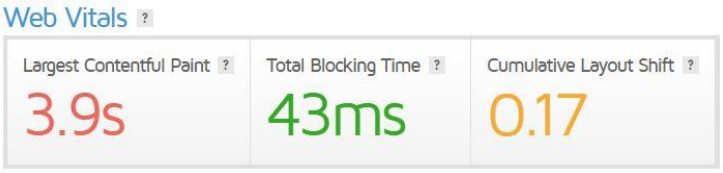


Рис. 1.9. Сторінка системи GTmetrix – швидкість надання доступу користувачам

Таким чином, у даному пункті обґрунтовується показник ефективності інформаційних веб-ресурсів, який стосується витрат часу на обробку запитів. Як випливає із вищенаведеного аналізу, існує спектр систем для оцінки ефективності веб ресурсів, проте усі вони використовуються для констатації факту ефективності того чи іншого веб-ресурсу. Разом з тим, жоден із них не дає можливості прогнозувати та моделювати показник ефективності. Вони лише придатні для збору поточної інформації про інформаційний веб-ресурс. Тому для використання показника ефективності необхідно розробити його математичне представлення і провести комп'ютерне моделювання та верифікацію запропонованого математичного опису з метою дослідження впливу структурованості контенту на ефективність веб-ресурсу. Такий підхід забезпечить не тільки встановлення цього впливу для конкретного веб-ресурсу,

але і забезпечить моделювання й оцінку структурованості контенту для розв’язування задач щодо його оптимізації, аналізу та перевірки на цілісність, достовірність й актуальність.

1.4. Постановка завдань дисертаційного дослідження

Спираючись на вище проведений аналіз методів, засобів та технологій організації інформаційних веб-ресурсів зробимо узагальнення та виділимо основні запитання для досліджень.

Розглянуто особливості організації інформаційних веб-ресурсів, показують, що якщо структура веб-сайту організована оптимально, то пошуковий робот витратить менше часу на аналіз усього ресурсу у тому числі і на аналіз контенту. Досліджено такі типи організації структури веб-сайтів: алфавітна організація структури; хронологічна; географічна; організація яка націлена на цільову аудиторію та тематична. Показано, що остання дозволяє швидко пересуватися веб-сайтом та оптимальна для SEO, оскільки веде від загальних понять до приватних концепцій і дозволяє реалізувати все семантичне ядро на сайті.

Ключовими вимогами до структури сайту є структура інформації на ресурсі, яка одночасно впливає на два показники: коефіцієнт користувача й ефективність, швидкість ранжування ресурсу пошуковою системою. Тому розробка схеми повинна одночасно відповідати вимогам як споживача, так і пошукової системи.

Проведено аналіз різних структур інформаційних веб-ресурсів та обґрунтовано актуальність дослідження організації веб-контенту на інформаційному порталі. Сформульовано основні показники для оцінювання веб-контенту та його релевантність запитам. Зазначено вимоги актуальностість та новинності та методи їх опису. Проведений аналіз показав, що ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити: продуктивність апаратного і програмного забезпечення, використаного джерелами, способи організації джерела, зокрема структурованість контенту та складність запиту.

Разом із тим, ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту, а також наявність у ньому застарілої та неактуальної інформації.

Наведено результати огляду сучасних методів та систем контент аналізу на основі веб-технологій.

Поняття контент-аналізу не має однозначного визначення, що породжує проблему: системи, побудовані на основі різних підходів до контент-аналізу, в загальному випадку несумісні. Узагальнено основні етапи дослідження контент-аналізу: виокремлення контенту; структурування; формалізація; реферування та аналіз із застосуванням методів математичної статистики та методів теорії ймовірності. Важливим питанням контент-аналізу є його автоматизація у спосіб використання спеціальних програмних систем.

У сучасних веб-технологіях реалізовано деякі процедури, які дають можливість інтегрувати системи контент-аналізу. Розглянуто особливості щодо реалізації методів контент-аналізу у технологіях Веб 1.0 – як мережа пізнання, веб 2.0 – як мережа спілкування, веб 3.0 – як мережа співпраці та веб 4.0 – як мережа інтеграції.

Веб 3.0 відомий як семантична мережа. Головною важливою метою семантичної мережі є зробити мережу читабельною для машин, а не тільки для людей. В цьому сенсі структурування та розпізнавання контенту є важливим завданням в межах зазначених технологій.

Разом із тим, формальний опис контенту, залишається актуальним, не залежно від стрімкого росту розглянутих технологій, а машинне навчання для його розпізнавання є ключовим процесом у виявлені недостовірної інформації, чи її неактуальності. Зважаючи на важливість процедур машинного навчання у контент-аналізі, виявлені недостовірної інформації, чи її неактуальності проведено аналіз існуючих методів машинного навчання. Встановлено, що генетичні алгоритми, подібно як і нейронні мережі вимагають формалізації представлення контенту та його кодування для використання в задачах автоматизованого аналізу контенту. Показано, що серед машинних методів навчання для структурування та виявлення некоректної, застарілої чи

неактуальної інформації найбільш легко адаптувати метод навчання на основі прецедентів.

Проведемо огляд відомих систем контент-аналізу. Системи контент-аналізу поділяються за типами: повністю автоматизовані пакети для контент-аналізу; надзвичайно прості та універсальні пакети підрахунку частот різних слів в текстах та напівавтоматичні пакети для ручного кодування текстових даних. До недоліків програм другого типу слід віднести відсутність категоризації підрахованих слів, що призводить до викривлення результатів в бік окремо беззмістовних слів, що належать до службових частин мови.

Огляд відомих систем для контент-аналізу із елементами машинного навчання продемонстрував:

1. Відсутність методу для моделювання базового показника оцінки ефективності функціонування веб-ресурсів залежно від структурованості контенту та оцінки ефективності функціонування під навантаженням.

2. Відсутність методів для формального опису контенту, що удосконалює процедури структурування, розпізнавання та порівняння текстового контенту веб-ресурсів.

Важливим питанням щодо удосконалення інформаційних веб-ресурсів, забезпечення їх ефективної роботи є вибір критеріїв оцінювання цієї ефективності. У процесі аналізу встановлено, що ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити продуктивність апаратного і програмного забезпечення, використовуваного джерелами, та, з іншого боку, способи організації джерела, зокрема структурованість контенту та складність запиту. Якщо перша група характеристик вирішується у спосіб перенесення ресурсу на програмно-технічні засоби з вищою продуктивністю, то друга група вимагає від проектувальників інформаційних веб-ресурсів зусиль, пов’язаних із забезпеченням високої якості проектування та оптимізації способів організації контенту. Обґрунтовується показник ефективності інформаційних веб-ресурсів, який стосується витрат часу на обробку запитів. Для використання цього показника необхідно розробити його математичне представлення і провести комп’ютерне моделювання та

верифікацію запропонованого математичного опису з метою дослідження впливу структурованості контенту на ефективність веб-ресурсу.

Таким чином, виходячи із вище викладеного актуальною є науково-прикладна задача створення програмної системи, яка забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів за рахунок застосування процедур структурування, розпізнавання і порівняння текстового контенту веб-ресурсів, виявлення у його структурі недостовірної, неточної чи неактуальної інформації.

Для розв’язання цієї задачі потрібно вирішити такі завдання:

1. Розглянути та обґрунтувати базові показники оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту.
2. Розробити новий метод моделювання показника оцінки ефективності функціонування інформаційних веб-ресурсів.
3. Розробити новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання та із використанням автоматизованого структурування контенту для виявлення недостовірної, неточної чи неактуальної інформації.
4. Вдосконалити методи для автоматизованого контент-аналізу, пошуку застарілої та некоректної інформації з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних.
5. Розробити архітектуру прикладної програмної системи, що виконує функції автоматизованого пошуку застарілої та некоректної інформації, яка містить модулі розпізнавання і порівняння текстового контенту веб-ресурсів із елементами машинного навчання.

Висновки до першого розділу

1. Розглянуто особливості організації та різні структури інформаційних веб-ресурсів та обґрунтовано актуальність дослідження організації веб-контенту на інформаційному порталі. Сформульовано основні показники для оцінювання веб-контенту та його релевантність запитам. Зазначено вимоги актуальності та новизни контенту та методи їх опису. Проведений аналіз показав, що ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити: продуктивність апаратного і програмного забезпечення, використаного джерелами, способи організації джерела, зокрема структурованість контенту та складність запиту. Разом з тим ефективність функціонування більшої кількості веб-ресурсів є невисока через слабку структурованість контенту, а також наявність у ньому застарілої та неактуальної інформації.

2. Проведено огляду сучасних методів та систем контент аналізу на основі веб-технологій. Узагальнено основні етапи дослідження контент-аналізу: виокремлення контенту; структурування; формалізація; реферування та аналіз із застосуванням методів математичної статистики та методів теорії ймовірностія, а також сформульовано вимоги до автоматизації цього процесу у спосіб використання спеціальних програмних систем. Встановлено, що у сучасних веб-технологіях реалізовано деякі процедури, які дають можливість інтегрувати системи контент-аналізу. Розглянуто особливості щодо реалізації методів контент-аналізу у технологіях Веб 1.0 – як мережа пізнання, веб 2.0 – як мережа спілкування, веб 3.0 – як мережа співпраці та веб 4.0 – як мережа інтеграції.

Показано, що формальний опис контенту, залишається актуальним, не залежно від стрімкого росту розглянутих технологій, а машинне навчання для його розпізнавання є ключовим процесом у виявлені недостовірної інформації, чи її неактуальності. Зважаючи на важливість процедур машинного навчання у контент-аналізі, виявлені недостовірної інформації, чи її неактуальності проведено аналіз існуючих методів машинного навчання. Встановлено, що генетичні алгоритми, подібно як і нейронні мережі вимагають формалізації представлення контенту та його кодування для використання в задачах

автоматизованого аналізу контенту. Показано, що серед машинних методів навчання для структурування та виявлення некоректної, застарілої чи неактуальної інформації найбільш легко адаптувати метод навчання на основі прецедентів.

3. Проведено огляд відомих систем контент-аналізу. Огляд відомих систем для контент-аналізу із елементами машинного навчання продемонстрував: відсутність методу для моделювання базового показника оцінки ефективності функціонування веб-ресурсів залежно від структурованості контенту та оцінки ефективності функціонування під навантаженням; відсутність методів для формального опису контенту, що удосконалює процедури структурування, розпізнавання та порівняння текстового контенту веб-ресурсів.

4. Важливим питанням щодо удосконалення інформаційних веб-ресурсів, забезпечення їх ефективної роботи є вибір критеріїв оцінювання цієї ефективності. У процесі аналізу встановлено, що ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити продуктивність апаратного і програмного забезпечення, використовуваного джерелами, та, з іншого боку, способи організації джерела, зокрема структурованість контенту та складність запиту. Обґрунтовано показник ефективності інформаційних веб-ресурсів, який стосується витрат часу на обробку запитів. Для використання цього показника необхідно розробити його математичне представлення і провести комп’ютерне моделювання та верифікацію запропонованого математичного опису з метою дослідження впливу структурованості контенту на ефективність веб-ресурсу.

На основі проведеного аналізу сформульовано завдання дисертаційного дослідження.

РОЗДІЛ 2
МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ЕФЕКТИВНОСТІ ФУНКЦІОНУВАННЯ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ В УМОВАХ РІЗНОЇ ЇХ СТРУКТУРОВАНОСТІ

Ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити продуктивність апаратного і програмного забезпечення, яке використовується джерелами, та способи організації джерела, зокрема структурованість контенту, і складність запиту. Якщо перша група характеристик вирішується способом перенесення ресурсу на програмно-технічні засоби з вищою продуктивністю, то друга група – вимагає від проектувальників інформаційних веб-ресурсів зусиль, пов'язаних із забезпеченням високої якості проектування та оптимізації способів організації контенту. Простіше кажучи – підвищення якості його структурування. У цьому контексті у даному розділі нами обґрунтовується показник ефективності інформаційних веб-ресурсів для витрат часу на обробку запитів. Для використання цього показника далі необхідно розробити його математичне представлення і провести комп'ютерне моделювання та верифікацію запропонованого математичного опису з метою дослідження впливу структурованості контенту на ефективність веб-ресурсу. Такий підхід забезпечить не тільки встановлення впливу для конкретного веб-ресурсу, але й забезпечить моделювання та оцінку структурованості контенту для розв'язання задач щодо його оптимізації, аналізу і перевірки на цілісність, достовірність та актуальність.

Таким чином, у цьому розділі розглядаються математичні моделі динаміки ефективності функціонування інформаційних веб-ресурсів. Обґрунтовано застосування стохастичних порогових моделей та запропоновано підхід до оцінювання параметрів цих моделей. Проведено чисельні експериментальні дослідження, що виконані на основі даних функціонування конкретного веб-ресурсу. На основі проведених експериментів підтверджено ефективність застосування запропонованих моделей.

Матеріали цього розділу ґрунтуються на опублікованих дисертантом працях [5-7].

2.1. Показник ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості

Для оцінки ефективності функціонування інформаційних веб-ресурсів потрібні математичні моделі процесів їх функціонування. Ці моделі дозволяють вирішувати завдання з оптимізації програмних алгоритмів і використання технічних ресурсів. Побудова таких моделей вимагає всебічного дослідження процесів функціонування веб-ресурсів із урахуванням впливу всіх елементів системи на її функціонування.

Процеси взаємодії з інформаційними джерелами полягають у формуванні запитів, перетворенні запитів у термінах джерел, отриманих та опрацьованих даних від джерел і формування результату, що за допомогою веб-сервера буде відправлений веб-браузеру. Це здійснюються підсистемою доступу до даних веб-ресурсу. Для взаємодії з джерелами використовується схема даних, яка містить опис всіх підключених до ресурсу джерел, опис типів даних, що реалізуються джерелами, і правила формування запитів до джерел у термінах цих джерел.

Для моделі ефективності інформаційних джерел як вихідний показник обрано величину часу, витрачену джерелом на виконання запитів на поточному інтервалі спостереження. Для моделі активності користувачів вихідним показником вибрано число запитів до інформаційних джерел на поточному інтервалі спостереження, при цьому зовнішні впливи виявляються поза полем уваги.

Оскільки для досліджуваних показників складно запропонувати моделі, побудовані на основі фізичних аналогій, передбачається їх ідентифікування. Характеристики розглянутих процесів демонструють поведінку нелінійних систем, оскільки для них характерні залежність збурень від поточних значень характеристики, циклічність, стрибкоподібні зміни.

Щоб урахувати нелінійну поведінку процесів і водночас мати можливість використовувати корисні властивості лінійних систем, пропонується підхід на основі класифікації можливих станів досліджуваної характеристики. Передбачається, що область значень характеристики може бути розбита на інтервали. Усередині інтервалу динаміка значень показника описується найпростішими лінійними рівняннями. При виході показника за межі інтервалу модель змінюється. Такий підхід запропонований у, де розглядається модель авторегресії з порогом (TAR, threshold autoregression).

Принципова відмінність моделі, що пропонується у дисертаційній роботі, від моделі TAR і TARMA полягає, перш за все, у тому, що значення модельованого показника виявляються доступні тільки частині, тому замість моделі часового ряду використовуються моделі динамічних систем спостереження. Крім того, передбачається можливість множинної зміни режимів у процесі еволюції системи спостереження.

Ефективність, тобто необхідні витрати часу на обробку запитів - це одна з основних характеристик інформаційного джерела. Як показують спостереження, час обробки запитів джерелами може змінюватися в широких межах. На цей процес впливає цілий набір неконтрольованих із боку ресурсу чинників, серед яких:

- різна продуктивність апаратного та програмного забезпечення, що використовується джерелами;

- різний обсяг даних джерела;

- різниця в обсязі результатів виконання запитів, пов’язана у тому числі зі складністю конкретного запиту;

- різна доступність джерел, пов’язана з характером роботи середовища.

Вплив перерахованих чинників важливий, проте їх вплив на кожне конкретне джерело у процесі його роботи змінюється незначно, тобто перераховані чинники можуть розглядатися як статичні. Однак є набір факторів, дія яких на джерело є не постійною, не прогнозованою, а саме:

- зміни, що вносяться до ресурсів джерела (додавання, видалення, модифікація екземплярів ресурсів);

- різний рівень навантаження на апаратне та системне програмне забезпечення з боку інших додатків, що існують на тій же платформі;

- різна завантаженість сегментів мережі для використання у взаємодії веб-ресурсу з джерелом;

- сформовані для користувача переваги при виборі переліку джерел, що беруть участь у виконанні запиту.

Таким чином, за показник ефективності інформаційних веб-ресурсів обираємо необхідні витрати часу на обробку запитів, а отримане його числове значення, залежно від різних чинників, використовуватимемо для математичної моделі.

При побудові моделі описаного явища повинні бути враховані наступні вимоги:

- сформована модель має враховувати залежність модельованого показника від часу (бути динамічною), оскільки, як було зазначено вище, на час виконання запиту впливає як інтенсивність, так і передісторія роботи з джерелом;

- при моделюванні має використовуватися дискретний час – інтервали спостереження повинні бути сформовані таким чином, щоб обчислення необхідної оцінки можна було здійснювати тільки в окремі моменти часу, розташовані один від одного на достатній відстані, з тим, щоб використані моделлю обчислювальні ресурси були досить малі порівняно з ресурсами, необхідними для забезпечення основної функціональності веб-ресурсу. Це необхідно для застосування моделі у процедурах оптимізації.

За рахунок зміни параметрів відповідних моделей стає можливим моделювання різних видів інформаційних джерел і режимів активності користувача. Як результат – з’являється можливість щодо аналізу працездатності й ефективності веб-системи з різними типами інформаційних джерел і режимами активності користувача при різних комбінаціях налаштувань.

2.2. Математична модель ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості

Виходячи із припущень щодо оцінки ефективності виконання запитів інформаційними ресурсами, замість кількісної оцінки ефективності пропонується ввести деякі умовні значення оцінки ефективності, визначивши категорії, за якими вони можуть бути згруповані. Ці категорії формуються із часових витрат на виконання запитів щодо певного виду інформаційного ресурсу.

Згрупуємо усі запити до інформаційних ресурсів за категоріями:

- σ_f – запити, що виконуються швидко;
- σ_m – запити, що виконуються із середньою швидкістю;
- σ_s – запити, що виконуються повільно.

Для кожної категорії модель спостережень можна подати у вигляді рівняння:

$$m_t = d_\sigma + e_\sigma \varphi_t \quad (1)$$

де m_t – середній час виконання запиту ресурсом за останній інтервал спостереження;

d_σ – середній очікуваний час виконання запиту ресурсом для поточної категорії σ запитів;

e_σ – значення, що визначає відхилення спостережуваних значень від середніх величин для поточної категорії запитів;

φ_t – випадкове збурення, що впливає на спостереження.

Для ресурсу S відомий показник r_t , що характеризує поточний режим його функціонування у певний момент часу t. Область значень R^1 показника r_t розіб'ємо на інтервали точками: $d: -\infty = d_0 < d_1 < \cdots d_{n-1} < d_n = +\infty$, де n – число режимів ефективності, підставивши у відповідність кожному режиму один інтервал. Режим, у якому знаходиться ресурс S, визначимо як приналежність r_t до цього інтервалу.

Вважаємо, що у будь-якому режимі середній час виконання запиту ресурсом за останній інтервал спостереження можна описати рівнянням (1), тобто, при зміні режиму m_t модель не змінюється, змінюються лише параметри. Останні можна замінити функціями $D(r_t)$ та $E(r_t)$, які повертають значення параметрів рівняння (1) залежно від поточного режиму, що визначається значенням показника r_t.

Виходячи з припущення, що залежність вимірюваних характеристик (час виконання запиту) для ресурсу S від його перебування в одному з режимів ефективності є лінійною, то можна запропонувати наступне рівняння для вимірювання у момент часу t характеристики:

$$m_t = D(r_t) + E(r_t)\varphi_t \tag{2}$$

Для $D(r_t)$ та $E(r_t)$ напишемо наступні вирази для їх зображення:

$$D(r_t) = \begin{Bmatrix} D_1, -\infty < x < d_1 \\ D_2, d_1 \le x < d_2 \\ \dots \\ D_n, d_{n-1} \le x < +\infty \end{Bmatrix} \tag{3}$$

$$E(r_t) = \begin{Bmatrix} E_1, -\infty < x < d_1 \\ E_2, d_1 \le x < d_2 \\ \dots \\ E_n, d_{n-1} \le x < +\infty \end{Bmatrix} \tag{4}$$

Значення $D_1, D_2, \dots, D_n$ та $E_1, E_2, \dots, E_n$ є відомими, виходячи із наведених вище припущень.

Для знаходження параметрів роботи веб-ресурсу, що описуються відношенням (2) розбиваємо інтервал спостережень $[t_0; +\infty]$ на деякі рівні відрізки $t_0 < t_1 < t_2 \dots t_{k-1} < t_k < \dots$. На поточному інтервалі $[t_{k-1}; t_k]$ для

кожного інформаційного ресурсу вимірюємо та зберігаємо час виконання відповідних запитів. У момент часу t_n формуємо із отриманих даних необхідну досліджену вибірку, що доповнюється середнім часом виконання запитів для кожного інформаційного ресурсу.

Довжина часового інтервалу $[t_{k-1}; t_k]$ повинна визначатися достатнім часом, щоб затрати обчислювальних ресурсів, які використовуються для оптимізації веб-ресурсу, були меншими, ніж затрати ресурсів, що витрачаються на роботу всього веб-ресурсу.

Для показника r_t повинна бути сформована модель, що описує закон, за яким ресурс S на поточному інтервалі спостереження $[t_{k-1}; t_k]$ має бути віднесений до тієї чи іншої категорії ефективності σ.

Зазначену модель, як і модель оцінки ефективності ресурсу, практично не можливо сконструювати виходячи з фізичного змісту процесів. Тому єдиним способом побудови таких моделей для стороннього спостерігача є застосування підходу на основі «чорної скриньки» із залученням методів параметричної ідентифікації [5].

Загальна рекомендація теорії ідентифікації моделей систем полягає ув тому, щоб починати ідентифікацію моделі на основі її простих структур [9]. Тому спочатку для побудови моделі обираємо проcту параметричну модель.

Припустимо, що для ресурсу S показник r_t є випадковим процесом, який описується рівнянням авторегресії першого порядку [7, 8]:

$$r_t = ar_{t-1} + s + b\psi_t \,, \tag{5}$$

де a, s, b - відомі невипадкові числа (задані, або такі, що підпадають під подальшу ідентифікацію), ψ_t - послідовність незалежних, однаково розподілених випадкових величин, r_0 - випадкова величина, що не залежить від ψ_t.

Для знаходження невідомих коефіцієнтів a, s, b авторегресії (5), на основі відомих експериментальних значень тривалості виконання запиту ресурсу, у

тому числі спостереження, використаємо середньоквадратичне відхилення між обчисленим значенням тривалості запиту та тим, що спостерігається, тобто:

$$\psi = \sum_{i=1}^{25}(ar_{t-1} + s + b\varphi_t - m_t)^2 \quad (6)$$

Тепер диференціюємо отриманий функціонал ψ (6) відносно невідомих коефіцієнтів a, s, b авторегресії і результати диференціювання квадратичної функції прирівняємо до нуля:

$$\begin{cases} \frac{\partial\psi}{\partial \mathrm{a}} = 0 \\ \frac{\partial\psi}{\partial \mathrm{s}} = 0 \\ \frac{\partial\psi}{\partial \mathrm{b}} = 0 \end{cases} \quad (7)$$

Отримана система (7) є системою лінійних алгебричних рівнянь. Її розв'язками є значення коефіцієнтів a, s, b авторегресії.

В якості об'єктивної числової характеристики T_t ефективності джерела S на момент часу t можна розглядати абсолютне значення часу виконання запитів ресурсом на поточному інтервалі спостереження $[t_{k-1}; t_k]$. Використання абсолютного значення T_t означає, що на його величину можуть впливати тільки ті чинники невизначеності, які безпосередньо пов'язані із цим ресурсом (складність виконаного запиту, обсяг даних ресурсу, технічні характеристики апаратного забезпечення і т.д.).

Для унеможливлення впливу різнорідності реальних джерел доцільно взяти відношення показника T_t до довжини інтервалу спостереження $[t_{k-1}; t_k]$, тобто перейти від абсолютного значення до відносного:

$$\mathrm{E}_q = \frac{T_{t_k}}{t_k - t_{k-1}}, \quad (8)$$

де E_q – показник відносної ефективності.

Досліджуючи динаміку цього показника, можна вважати, що

$$r_{t_k} = r_{t_{k-1}} + \frac{T_{t_k}}{t_k - t_{k-1}} \tag{9}$$

Використавши для опису r_t модель, що представлена відношенням (2), залишається вибрати параметри, що забезпечують адекватність моделі спостереження до реально отриманих даних.

Наявність процесу r_t пов'язано з необхідністю моделювати поточний режим ефективності інформаційного ресурсу. Даний процес необхідний, щоб забезпечити послідовність зміни режимів, що відповідає спостережуваним даним. Зміна режиму відбувається при виході значень процесу за межі інтервалу, визначеного для обраного режиму. У даному контексті величина та розмірність показника r_t не має значення.

В якості моделі, яка описує динаміку показника ефективності веб-ресурсу, пропонується наступна стохастична динамічна система спостережень загального вигляду:

$$\begin{cases} r_t = a_t r_{t-1} + s_t + b_t \psi_t, \ t = 1,2,\dots, \\ m_t = D(r_t) + E(r_t)\varphi_t \end{cases}, \tag{10}$$

де φ_t не залежить від ψ_t, r_0.

Таким чином, є можливість обґрунтованого вибору параметрів розглянутої моделі за рахунок нескладного статистичного аналізу середовища функціонування вибраного інформаційного ресурсу.

2.3. Методика отримання експериментальних даних

Як було зазначено вище, для побудови математичної моделі ефективності функціонування інформаційних веб-ресурсів в умовах різної їх структурованості, необхідно розв'язати задачу ідентифікації її параметрів. Процедура ідентифікації параметрів вимагає використання результатів

експерименту для конкретного інформаційного веб-ресурсу. Очевидно, що у цьому випадку досліднику необхідно мати доступ до вихідного коду.

У нашій дисертаційній роботі запропонована методика отримання експериментальних даних з веб-ресурсу та конкретні алгоритми, представлені за допомогою скриптів.

Для отримання даних про потрібний контент та збір статистики на веб-ресурсі застосовується Google API.

Для цього створюється файл time _ query.php із таким кодом:

```
<?php
require_once __DIR__ . '/vendor/autoload.php';
$analytics = initializeAnalytics();
$response = getReport($analytics);
printResults($response);
function initializeAnalytics()
```

Згенерований ключ містить ім’я JSON-файла:

```
{
  $KEY_FILE_LOCATION = __DIR__ . '/service-account-
credentials.json';
  $client = new Google_Client();
  $client->setApplicationName("Top pages");
  $client->setAuthConfig($KEY_FILE_LOCATION);
  $client-
>setScopes(['https://www.googleapis.com/auth/analytics.readonly'])
;
  $analytics = new Google_Service_AnalyticsReporting($client);
  return $analytics;
}
```

Значення свого VIEW_ID:

```
$VIEW_ID = "<REPLACE_WITH_VIEW_ID>";
```

DateRange – створюється об’єкт (діапазон дат).

```
      $dateRange = new
Google_Service_AnalyticsReporting_DateRange();
```

Встановлення початкової та кінцевої дати вибірки:

```
$dateRange->setStartDate("12daysAgo");
 $dateRange->setEndDate("yesterday");
```

Для збору статистики часу виконання запитів застосовується Core Reporting API.

Потрібно виконати наступні дії:

1. Надіслати запит до Core Reporting API.
2. Перевірити властивість containsSampledData відповіді, що визначає наявність вибірки.
3. Якщо вибірка має значення true, можна зробити запит на детальний звіт за допомогою полів query і profileInfo.

Приклад поля query у відповіді Core Reporting API:

```
"query": {
  "start-date": "2020-01-01",
  "end-date": "2020-01-31",
  "ids": "ga:1234",
  "dimensions": "ga:browser",
  "metrics": [
  "ga:visits"
  ],
  "filters": "ga:country==US",
  "start-index": 1,
  "max-results": 1000
}
```

Приклад поля profileInfo у відповіді Core Reporting API:

```
"profileInfo": {
  "profileId": "1234",
  "accountId": "12345",
  "webPropertyId": "UA-12345-1",
```

```
  "internalWebPropertyId": "11254",
  "profileName": "Name of the profile",
  "tableId": "ga:1234"
}
```

Детальний звіт за допомогою Core Reporting API:

```
// Make a Core Reporting API call.
GaData reportingApiData = v3.data().ga().get(...).execute();

// Check if the response is sampled.
if (reportingApiData.getContainsSampledData()) {

  // Use the “query” object to construct an unsampled report
object.
  Query query = reportingApiData.getQuery();
  UnsampledReport report = new UnsampledReport()
      .setDimensions(query.getDimensions())
      .setMetrics(Joiner.on(',').join(query.getMetrics()))
      .setStartDate(startDate)
      .setEndDate(endDate)
      .setSegment(query.getSegment())
      .setFilters(query.getFilters())
      .setTitle(“My unsampled report”);

  // Use “profileInfo” to create an InsertRequest for creating
an
  // unsampled report.
  ProfileInfo profileInfo = reportingApiData.getProfileInfo();
  Insert insertRequest =
analytics.management().unsampledReports()
  .insert(profileInfo.getAccountId(),
          profileInfo.getWebPropertyId(),
          profileInfo.getProfileId(),
          report);
  UnsampledReport createdReport = insertRequest.execute();
}
```

Якщо ж запит без вибірки повертає великий обсяг даних, потрібно встановити:

1. Діапазон дат (початкова дата – старт проекту; кінцева дата – поточна дата);
2. Показник мети, що дорівнює ga: pagetimes;
3. Параметр ga: pagePath, значення якого має відповідати URL певної сторінки.

Діапазон дат:

```
$dateRange = new
Google_Service_AnalyticsReporting_DateRange();

$dateRange->setStartDate("2020-01-01");

$dateRange->setEndDate("today");
```

Показник мети, що дорівнює ga: pagetimes;

```
$pagetimes = new Google_Service_AnalyticsReporting_Metric();

$pagetimes->setExpression("ga:pagetimes");

$pagetimes->setAlias("pagetimes");
```

Створення фільтру за параметром ga: pagePath

```
$dimensionFilter = new
Google_Service_AnalyticsReporting_DimensionFilter();

$dimensionFilter->setDimensionName('ga:pagePath');

$dimensionFilter->setOperator('EXACT');

$dimensionFilter->setExpressions('/html-and-css/css-
selector');
```

Створення групи фільтрів за параметром:

```
$dimensionFilterClause = new
Google_Service_AnalyticsReporting_DimensionFilterClause();

$dimensionFilterClause->setFilters($dimensionFilter);
```

Створення об’єкта ReportRequest (запит) :

```
$request = new
Google_Service_AnalyticsReporting_ReportRequest();
```

Ідентифікатор запиту viewId :

```
$request->setViewId($VIEW_ID);
```

Діапазон дат:

```
$request->setDateRanges($dateRange);
```

Показники запиту:

```
$request->setMetrics(array($pagetimes));
```

Фільтри запиту:

```
$request->setDimensionFilterClauses([$dimensionFilterClause]);
```

Отримані дані можна вивантажити у форматі JSON в певний файл. Потрібно написати відповідну функцію, наприклад, printResultInFile, і використати її для обробки відповіді:

```
printResultInFile($response);
```

Для вимірювання середнього час відгуку веб-ресурсу написано скрипт:

```
<?php
function array_compare($time1, $time2)
{
    if (count($time1) < count($time2)) {
        return -1; // $time1 < $time2
    } elseif (count($time1) > count($time2)) {
        return 1; // $time1 > $time2
    }
    foreach ($time1 as $key => $val) {
        if (!array_key_exists($key, $time2)) {
```

```
                return null; // не можуть бути порівнянні
            } elseif ($val < $time2[$key]) {
                return -1;
            } elseif ($val > $time2[$key]) {
                return 1;
            }
        }
        return 0; // $time1 == $time2
    }
    ?>
```

Таким чином, представлена методика отримання експериментальних даних про ефективність інформаційного веб-ресурсу та алгоритми виконання конкретних дій, створюють можливість ефективного збору та оцінки функціонування будь-якого веб-ресурсу з доступним вихідним кодом. Результати апробації запропонованої методики проведено на конкретному веб-ресурсі та наведено у наступному підрозділі.

2.4. Приклад ідентифікації математичної моделі ефективності функціонування інформаційних веб-ресурсів

Застосування описаної вище процедури розглянемо на прикладі веб-ресурсу кафедри комп’ютерних наук Тернопільського національного економічного університету. Було проведено збір статистики часу виконання запитів трьома інформаційними джерелами: інтегрованою системою інформаційних ресурсів (IIS), персональними сторінками викладачів кафедри (PPT) і системою зберігання неструктурованого контенту цього веб-ресурсу (UCWR). Для кожного джерела необхідно виміряти середній час відповіді в секундах із часовим інтервалом 2 хвилини.

На рисунку 2.1 представлено фрагменти вибірки даних для кожного інформаційного джерела (IIS), (PPT), (UCWR).

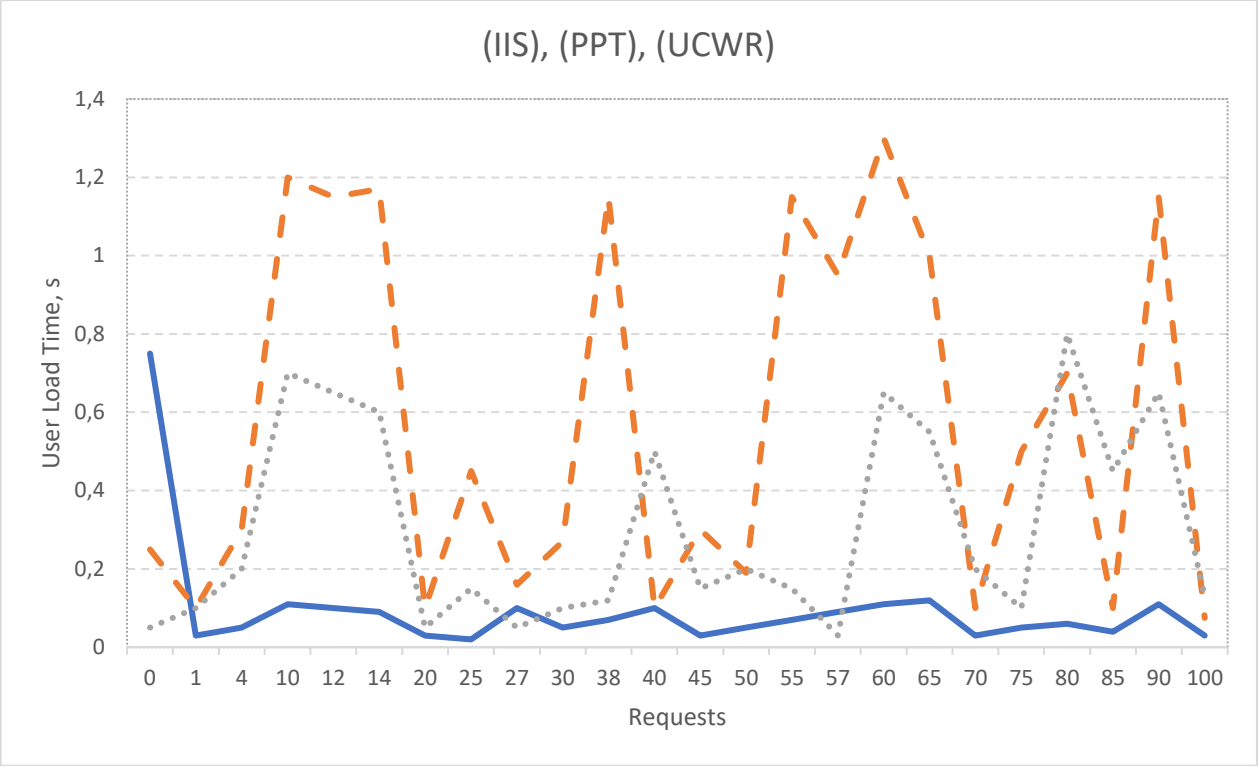


Рис. 2.1. Експериментальні дані для кожного інформаційного джерела (IIS), (PPT), (UCWR) ресурсу http://www.dcs.tneu.edu.ua/

На рисунку 2.1 видно, що тривалість виконання запитів розподілено нерівномірно, існують деякі значення, біля яких знаходиться велика частина спостережень. Особливо великі часові затримки для спостережень, що характеризують зберігання неструктурованого контенту досліджуваного веб-ресурсу. Результати вимірювання тривалості виконання запиту зводили до середнього для серій з декількох вимірювань (чотирьох). Результати усереднених вимірювань тривалості виконання запитів у розрізі категорій, наведено у таблиці 2.1.

Таблиця 2.1.

Результати вимірювання ефективності джерел із різним рівнем структурованості контенту (IIS) (PPT), (UCWR) ресурсу http://www.dcs.tneu.edu.ua/

№	IIS (інтегрована система інформаційних ресурсів)	PPT (персональні сторінки викладачів кафедри)	UCWR (система зберігання неструктурованого контенту веб-ресурсу)
Серії експерименту	r_{ft}	r_{mt}	r_{st}
1	0.8	0.1	0.3
2	0	0.1	0.1
3	0.1	0.2	0.3
4	0.1	0.7	1.2
5	0.1	0.7	1.2
6	0.1	0.6	0.1
7	0	0.1	0.5
8	0	0.2	0.2
9	0.1	0.1	0.3
10	0.1	0.1	1.2
11	0.1	0.1	0.1
12	0.1	0.5	0.3
13	0	0.2	0.2
14	0.1	0.2	1.2
15	0.1	0.2	1
16	0.1	0	1.3
17	0.1	0.7	1
18	0.1	0.6	0.1

Продовження таблиці 2.1.

19	0	0.2	0.5
20	0.1	0.1	0.4
21	0.1	0.8	0.1
22	0	0.5	1.2
23	0.1	0.7	0.1
24	0.2	0.1	0.6
25	0.1	0.3	0.7

Середньоквадратичні значення є результатом обчислень на основі даних, наведених у таблиці 2.1 для кожної категорії інформаційного веб-ресурсу. Тобто, середні та середньоквадратичні відхилення тривалості запитів для кожної категорії є результатом, отриманим на основі 25 спостережень експерименту для кожної категорії.

Усі запити до інформаційних ресурсів розділені за трьома категоріями: $D_f, D_m D_s$ - середній очікуваний час виконання запиту ресурсом для поточної категорії σ запитів. $E_f, E_m E_s$ – значення, що визначає відхилення спостережуваних значень від середніх величин для поточної категорії запитів.

Параметри $D_f, D_m D_s, E_f, E_m E_s$ визначалися як середній час виконання запиту відповідним ресурсом, який відносився до певної категорії, та середньоквадратичним відхиленням, що характерний для ресурсу відповідної категорії, відповідно.

Результати оцінювання параметрів $D_f, D_m D_s\ E_f, E_m E_s$ наведено у таблиці 2.2.

Таблиця 2.2.

Результати обчислення середніх значень та середньоквадратичних відхилень тривалості запитів для кожної категорії з різним рівнем структурованості контенту (IIS) (PPT), (UCWR) ресурсу http://www.dcs.tneu.edu.ua/

Середнє значення тривалості запитів для IIS (інтегрованої системи інформаційних ресурсів)	Середнє значення тривалості запитів для PPT (персональних сторінок викладачів кафедри)	Середнє значення тривалості запитів для UCWR (система зберігання неструктурованого контенту веб-ресурсу)
D_f	D_m	D_s
0,104	0,32	0,564
Середньоквадратичне відхилення тривалості запитів для IIS (інтегрованої системи інформаційних ресурсів)	Середньоквадратичне відхилення тривалості запитів для PPT (персональних сторінок викладачів кафедри)	Середньоквадратичне відхилення тривалості запитів для UCWR (системи зберігання неструктурованого контенту веб-ресурсу)
E_f	E_m	E_s
0,151	0,260	0,458

Значення із таблиці 2.1 були використані для визначення відповідних категорій ефективності. До категорії σ_f віднесено спостереження зі значенням близько 0.1 секунди, до категорії σ_s - спостереження зі значенням близько 1 секунди. До категорії σ_m віднесені спостереження в інтервалі між 0.1 та 1 секундою. Далі інтервали були скориговано так, щоб врахувати збурення значень. Ці інтервали також зроблено такмими, що не перетинаються, щоб виключити частину значень, які знаходяться поблизу меж інтервалів і тому не можуть бути однозначно віднесені до тієї чи іншої категорії.

Таким чином, для отриманих даних визначено три інтервали $\Delta_f^r = (0; 0.15]$, $\Delta_m^r = (0.2; 0.45]$, $\Delta_s^r = (0.56; +\infty)$. Із вибірки для кожного ресурсу сформовано три множини значень, відповідно до обраних інтервалів. Далі оцінено ймовірності віднесення кожного ресурсу до відповідної категорії.

Ймовірності визначалися частотами попадання спостережуваних даних у відповідні інтервали.

Таблиця 2.3.

Оцінка ймовірності приналежності ресурсу до відповідної категорії ефективності

Ресурс	$P(r_t \in \Delta_f)$	$P(r_t \in \Delta_m)$	$P(r_t \in \Delta_s)$
IIS	0.95	0.03	0.02
PPT	0.74	0.19	0.07
UCWR	0.28	0.37	0.33

Для визначення параметрів a, s, b процесу (5) в якості випадкових збурень використано стандартний гаусівський білий шум, а інтервали $\Delta_f^r = (-\infty, -10]$, $\Delta_m^r = (-10,10]$, Δ_s^r=$(10, +\infty]$.

У цих припущеннях параметри нормального розподілу визначено шляхом вирішення системи рівнянь, що отримано із визначення властивостей щільності ймовірності.

Безпосередньо зі спостережень не можна зробити висновок про визначення параметра a (для малої початкової кількості вимірювань), тому проведено кілька експериментів із моделювання ефективності ресурсів із різними значеннями параметра a. Із отриманих результатів вибрано значення параметра a, що презентує найбільш близьку поведінку моделі порівняно з фактичними значеннями. На пізніших стадіях, коли кількість вимірювань тривалості запиту була більшою від кількості коефіцієнтів авторегресії, була використано методику, наведену у попередньому підрозділі (мінімізації середньоквадратичного відхилення виміряного та прогнозованого значень).

Оцінки значення параметра a й обчислених значень параметрів s,b наведено у таблиці 2.4.

Таблиця 2.4.

Оцінки параметрів a, s, b процесу (5)
для відповідної категорії ефективності

Ресурс	A	s	b
IIS	0.1	-83	50.5
PPT	0.2	-20	23.6
UCWR	0.3	0.73	19.9

На основі знайдених показників a, s, b – заздалегідь відомі невипадкові числа (задані, або такі, що передбачають подальшу ідентифікацію у процесі вибірки), що впливають на показник r_t, який є випадковим процесом, що описується рівнянням авторегресії першого порядку (5) для досліджуваного ресурсу S. Дані показники впливають на функції $D(r_t)$ та $E(r_t)$, які повертають значення параметрів рівняння (1) залежно від поточного режиму, що визначається значенням показника r_t.

Результат моделювання для системи зберігання неструктурованого контенту аналізованого веб-ресурсу представлено на рисунку 2.2.

Отримані моделі забезпечують значення статистичних параметрів (середнє значення і середньоквадратичне відхилення) кожної категорії ефективності ресурсів і частот віднесення ресурсу до категорії, що близькі до параметрів спостережень.

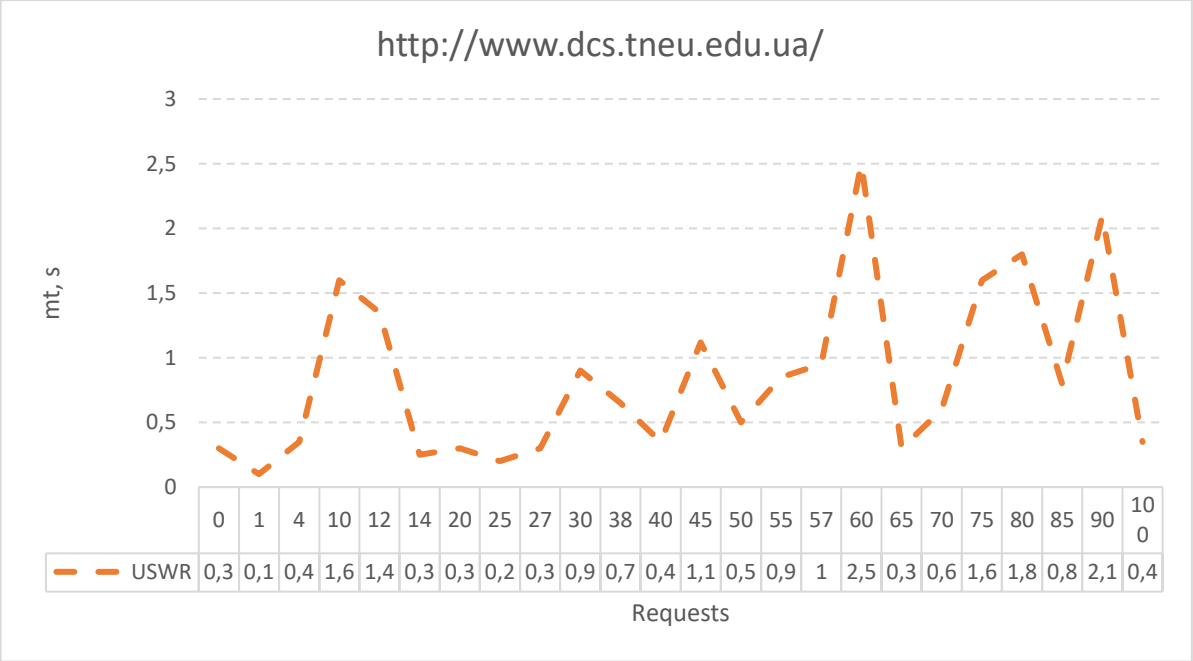


Рис. 2.2. Результати моделювання для неструктурованого контенту ресурсу http://www.dcs.tneu.edu.ua/

Слід зауважити, що таблиця 2.3 відображає результати, що засвідчують найчастіше звертання користувача, а саме до добре структурованого веб-ресурсу. Зауважимо, що у цьому випадку імовірності належності до діапазону тривалості виконання запитів добре структурованого контенту для різних категорій контенту розподілилися таким чином: $P(r_t \in \Delta_f)$ ={0.95; 0.74; 0.28}.

В інших випадках це б показувало службі підтримки потребу у структуруванні контенту, бо проміжок тривалості виконання запитів був би віднесений до інших категорій: другої $r_t \in \Delta_m$та третьої $r_t \in \Delta_s$ (див. табл. 3). Тобто це свідчило б про низьку ефективність інформаційного веб-ресурсу.

Висновки до другого розділу

1. Запропоновано та обґрунтовано підхід до моделювання базових показників оцінки ефективності функціонування інформаційних веб-ресурсів. Він дозволяє врахувати характерні для нелінійних систем особливості, такі як: залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристик. Для врахування таких явищ, з одного боку, і для можливості використання корисних властивостей лінійних систем ідентифікації, з іншого, у роботі реалізується підхід на основі класифікації можливих станів досліджуваного показника. В області значень показника вибираються інтервали. При знаходженні значень показника всередині інтервалу його динаміка описується найпростішими лінійними рівняннями.

2. Встановлено можливість обґрунтованого вибору параметрів розглянутих моделей, тобто їх ідентифікацію на основі експериментальних даних, за рахунок нескладного статистичного аналізу середовища функціонування веб-ресурсу. Описано методику отримання експериментальних даних із інформаційного веб-ресурсу та розроблено алгоритмічні і програмні засоби для її реалізації.

3. Наведено приклад застосування процедур оцінювання параметрів моделі на основі експериментальних даних, зібраних у процесі функціонування сайту кафедри комп’ютерних наук Тернопільського національного економічного університету www.dcs.tneu.edu.ua.

4. Отримані на основі запропонованого підходу математичні моделі можуть бути використані при вирішенні цілого ряду практичних завдань, що виникають при розробці і впровадженні веб-систем: налаштування параметрів продуктивності, аналіз впливу структурованості контенту, оцінка ефективності функціонування під навантаженням.

РОЗДІЛ 3
МЕТОДИ ТА ЗАСОБИ ВИЯВЛЕННЯ НЕКОРЕКТНОЇ ТА НЕАКТУАЛЬНОЇ ІНФОРМАЦІЇ НА ОСНОВІ КОНТЕНТ-АНАЛІЗУ ВЕБ-РЕСУРСІВ

Як зазначено у першому розділі, для усунення некоректної та неактуальної інформації на веб-ресурсах, зокрема на інформаційних веб-ресурсах, широко використовують методи контент-аналізу. Переважно такі функції виконують програмні системи автоматично. Основна ідея побудови системи полягає у перетворенні інформації із інформаційних веб-ресурсів у формат, придатний для порівняння із базами даних, що існують в організаціях та порівняння їх. У цьому розділі нами розглянуто саме такий підхід.

Його застосування апробовано автором для автоматизованого пошуку застарілої та некоректної інформації на тих сайтах закладів вищої освіти, де було отримано доступ до відповідних баз даних. Однак не завжди є таке можливим. Тому у дисертаційній роботі рекомендовано ще один метод виявлення некоректної та неактуальної інформації на основі контент-аналізу веб-ресурсів, що побудовано на принципах мажоритарного голосування (за аналогією з такими методами у теорії надійності систем).

У завершальній частині розділу запропоновано та обґрунтовано метод виявлення некоректного та недостовірного контенту із використанням засобів машинного навчання.

Результати досліджень, що презентовані у цьому розділі, опубліковано дисертантом автором у таких працях [1-3].

3.1. Методи та засоби виявлення некоректної та неактуальної інформації на основі існуючих баз даних контенту

Задачі аналізу контенту і виявлення некоректної та застарілої інформації є достатньо складними. Для їх розв’язання потрібно створювати інтелектуальні системи із моделюванням контенту на основі використання онтологічного підходу. Проте на початкових етапах необхідно встановити певні особливості як

існуючих алгоритмів аналізу контенту, так і майбутніх – націлених не тільки на звичайний контент, а й на виявлення, наприклад, неактуальної чи недостовірної інформації.

Система аналізу контенту повинна складатися з двох частин. Перша – встановлює семантичну структуру сайту, спираючись на конкретний код, наприклад на HTML, проводити семантичний аналіз (парсинг) і запис результатів у базу даних. Друга частина повинна із аналізу доступних веб-ресурсів з подібним контентом проводити семантичний аналіз (парсинг) і сформувати аналогічну базу даних для зіставлення із даними бази, сформованої першою частиною. Функції першої системи опрацьовано в ряді праць [3]. Зокрема, побудовано моделі представлення семантичної структури веб-сайтів за умови доступу до програмного коду сторінки.

Більшість веб-сторінок створюють за допомогою мови HTML (англ. HyperText Markup Language – мова розмітки гіпертекстових документів) або XHTML. Браузер опрацьовує документ HTML та відтворює на екрані. Отже, усі веб-сайти мають однакову HTML-розмітку, що дає можливість репродуктувати HTML-документ за допомогою браузера у звичному для користувача вигляді. Власне, контент розміщують у блоці <body>, але також у цьому блоці теги можуть містити різні імена та послідовність. Останнє означає певні частини контенту, що суттєвим чином ускладнює його аналіз.

Наступною функцією системи є парсинг сайтів – послідовний синтаксичний аналіз інформації, розміщеної на інтернет-сторінках [4]. Текст інтернет-сторінок – ієрархічній набір даних, структурований за допомогою природних і комп’ютерних мов. Веб-парсинг використовуємо для автоматизованого збору контенту або даних із будь-якого сайту чи сервісу. Результат парсингу записуємо у базу даних попередньо встановленої структури.

Спираючись на аналіз існуючих систем контент-аналізу веб-сайтів, можемо сформулювати основні вимоги до алгоритмів автоматизованого виявлення застарілої та недостовірної інформації.

Серед переліку вимог: налаштування фільтрів URL, щоб не парсити зайві сторінки, налаштування глибини парсингу, якісне скачування контенту, обробка у декілька потоків та збереження контенту у різних форматах (Рис.3.1).

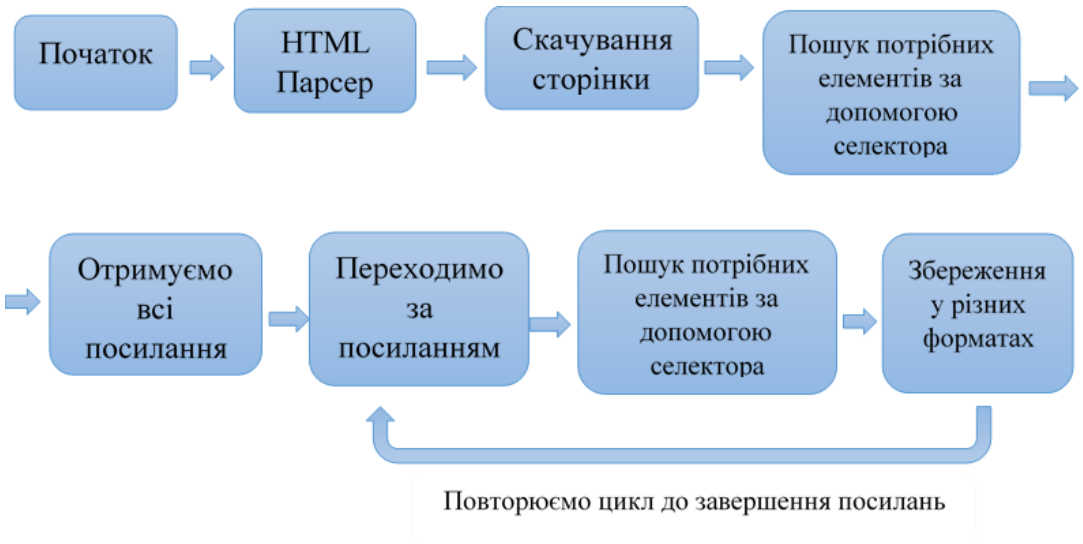


Рис. 3.1. Схематичне зображення збору інформації за допомогою парсингу та збереження інформації

Як бачимо з рис. 3.1, після здійснення парсингу згенеровані результати необхідно опрацювати, щоб надати інформації вигляд, придатний для подальшого використання. Конкретний формат залежить від того, як у подальшому будуть оброблятися зібрані дані. Доволі часто з парсенного контенту за допомогою XML формується RSS-потік, що зручний для використання даних без процедури рерайтингу. Іноді результат парсингу розміщують у CSV-файл, оскільки цей текстовий формат дуже простий у подальшій обробці, легко конвертується у SQL-запити і без проблем відкривається в Excel. В особливих випадках потрібно, щоб кінцеві дані були представлені у вигляді електронних таблиць XLS. Складнішим завданням, є реалізація функцій другої частини системи виявлення некоректної та неактуальної інформації на основі контент-аналізу веб-ресурсів. У нашій дисертаційній роботі для реалізації системи запропоновано використати порівняння результатів парсингу із відомою базою організації - власника інформаційного веб-ресурсу.

Загальну схему реалізації методу виявлення некоректної та застарілої інформації на веб-ресурсах і порівняння із існуючою базою даних наведенао на рис. 3.2.

Як бачимо із нижченаведеної на рис. 3.2. архітектури, процес підготовки контенту до процедури виявлення неактуальної та недостовірної інформації на веб-ресурсах передбачає виконання таких дій:

1. Збір контенту, найчастіше шляхом скачування веб-сторінки.
2. Отримані дані HTML-формату трансформуються у заданий формат.
3. Перетворення отриманих результатів у зручний для користувача та доступний файл, наприклад, формування бази даних.

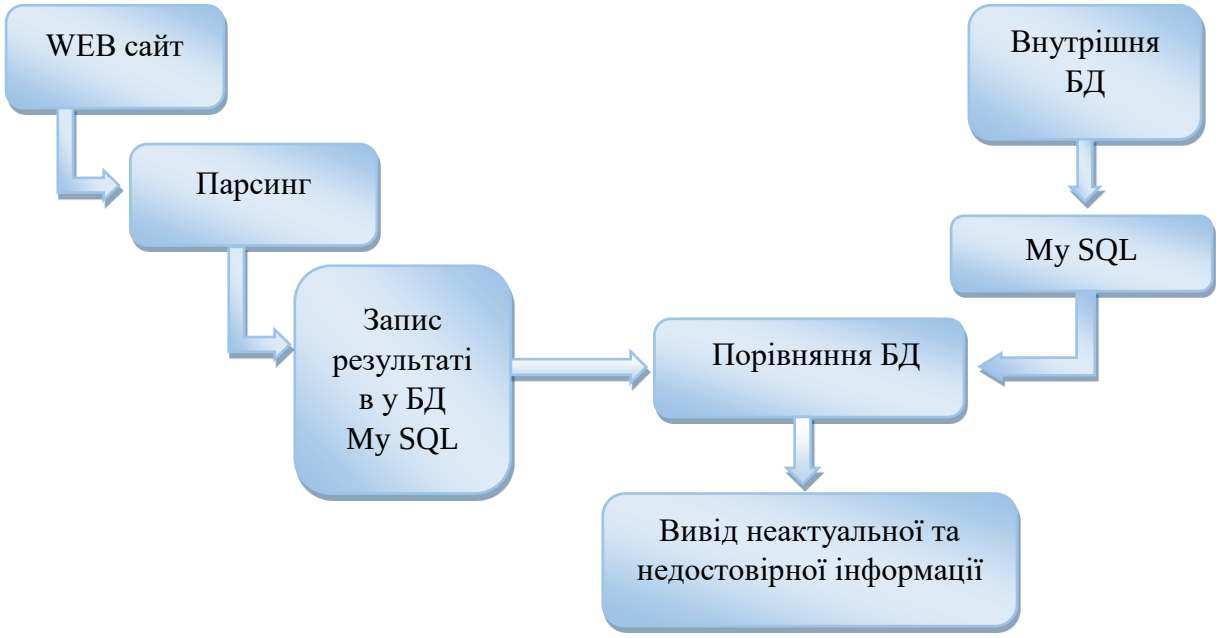


Рис. 3.2. Схема реалізації методу виявлення неактуальної та недостовірної інформації на веб-ресурсах і порівняння із існуючою базою даних

Після отримання результатів та запису у базу даних потрібно їх синхронізувати із еталонною інформацією.

Для початку синхронізації даних у БД варто отримати перелік усіх елементів, які співставляються.

Скористаємося запитом, що порівнює стовпці (імена) у таблицях між двома базами даних (схемами) MySQL.

```
set @database_1 = 'real_object';
set @database_2 ='rezult';
```

Основним взастосуванням команди SELECT є перегляд рядків з таблиці. Нижче наведено приклад команди SELECT, що відобразить усі рядки з таблиці:

```
select *
from (
        select COALESCE(c1.table_ rezult,
c2.table_real_object) as table_name,
               COALESCE(c1.column_name, c2.column_name) as
table_column,
               c1.column_name as database1,
               c2.column_name as database2
        from
            (select table_ rezult,
                    column_name
             from information_schema.columns c
             where c.table_schema = @database_1) c1
        right join
                (select table_ real_object,
                        column_name
                 from information_schema.columns c
                 where c.table_schema = @database_2) c2
        on c1.table_name = c2.table_name and c1.column_name =
c2.column_name
     union
```

Функція COALESCE приймає ряд аргументів і повертає перший аргумент, що не має значення NULL. Якщо всі аргументи NULL, функція COALESCE повертає NULL.

Нижче наведено приклад використання функції COALESCE:

```
        select COALESCE(c1.table_ rezult e, c2.table_ real_object)
as table_name,
               COALESCE(c1.column_name, c2.column_name) as
table_column,
                   c1.column_name as schema1,
                   c2.column_name as schema2
            from
                (select table_ rezult,
                       column_name
                 from information_schema.columns c
                 where c.table_schema = @database_1) c1
            left join
                     (select table_ real_object,
                            column_name
                      from information_schema.columns c
                      where c.table_schema = @database_2) c2
            on c1.table_name = c2.table_name and c1.column_name =
c2.column_name
     ) tmp
     where database1 is null
           or database2 is null
     order by table_ rezult,
              table_column;
     set @database_1 = null;
     set @database_2 = null;
```

3.2. Метод та засоби виявлення некоректної та неактуальної інформації із застосуванням принципу “мажоритарного голосування”

Складнішим завданням, порівняно із викладеним у попередньому підрозділі, є реалізація функцій другої частини системи виявлення некоректної та неактуальної інформації на основі контент-аналізу веб-ресурсів без наявної бази даних.

Розглянемо реалізацію запропонованого формального підходу із додаванням процедур визначення коректної та актуальної інформації на основі аналізу існуючих веб-ресурсів.

У процесі контент-аналізу з використанням пошукових веб-браузерів отримано n веб-ресурсів, що містять контент, подібний до контенту аналізованого інформаційним веб-ресурсом.

Представимо структуру цієї інформації за допомогою схеми на рисунку 3.3.

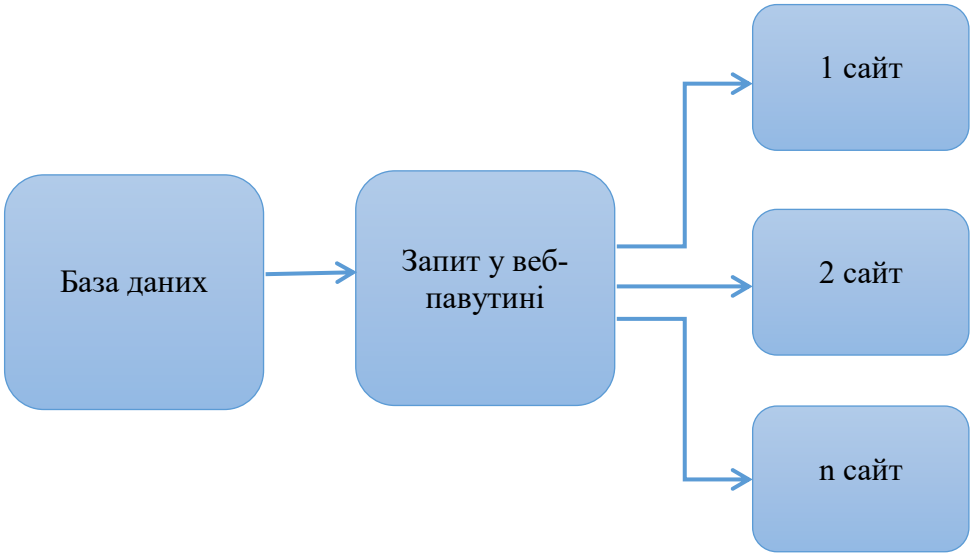


Рис. 3.3. Структура запиту та результатів пошуку на трьох веб-ресурсах

Тепер за принципом “мажоритарного голосування”, його різних можливостей порівняємо цей контент. Наприклад, «2 із 3», як це показано на рис. 3.4.

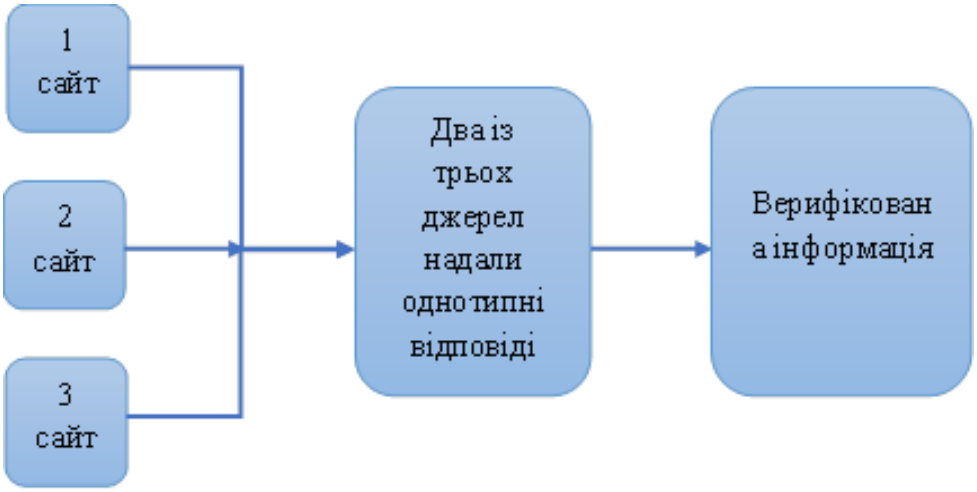


Рис. 3.4. Схема реалізації методу “мажоритарного голосування” із використанням трьох джерел

На рис. 3.5. наведено схему реалізації методу виявлення некоректної чи неактуальної інформації за принципом “мажоритарного голосування” «3 із 5».

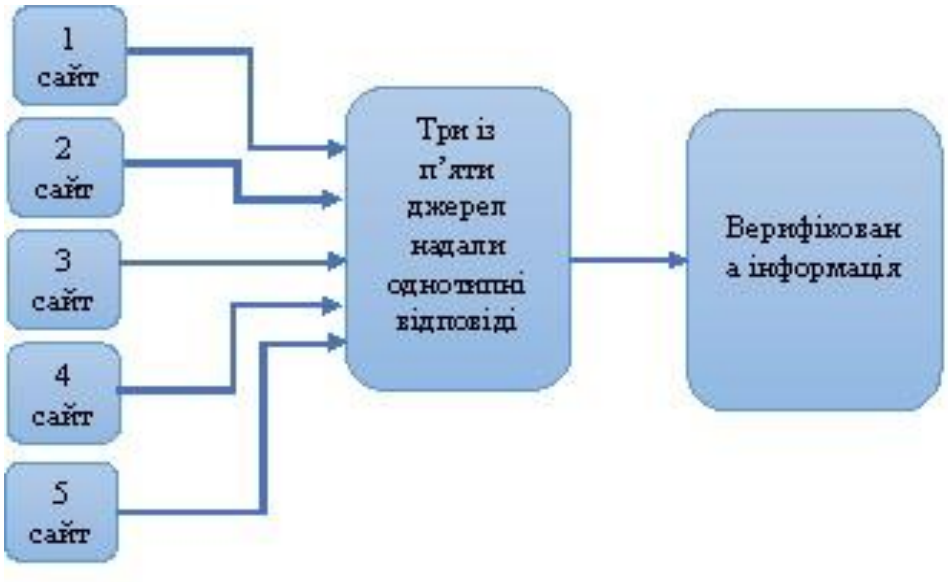


Рис. 3.5. Схема реалізації методу “мажоритарного голосування” із використанням п’яти джерел

У результаті застосування принципу “мажоритарного голосування” встановлюємо подібність структурованої інформації на різних веб-ресурсах. Це означає, що якщо на 3 ресурсах було встановлено подібний контент, який стосується деякого об’єкту, і 2 з них надали контент, який співпадає, тоді приймаємо рішення, що він коректний та актуальний (рис. 3.4.). Якщо ми використовуємо принцип “мажоритарного голосування” «3 з 5», і на 5 ресурсах було встановлено подібний контент, який стосується деякого об’єкту, та 3 із них надали контент, який співпадає, тоді приймаємо рішення, що він є коректним та актуальним теж (рис. 3.5.).

Приймаючи гіпотезу, що інформація повторюється при різних комбінаціях веб-ресурсів, які не пов’язані між собою, то це означає, що інформація є коректною та актуальною.

Таким чином, реалізація принципу “мажоритарного голосування” дає можливість підвищити коректність та актуальність інформації, отриманої при контент-аналізі.

Для розробки системи автоматичного керування контентом використовуємо семантичну структуру інформаційного веб-ресурсу. Алгоритм побудови семантичного аналізу ґрунтується на аналізі структури веб-сторінки.

Веб-сторінок створюють за допомогою мови HTML (англ. HyperText Markup Language – мова розмітки гіпертекстових документів) або XHTML. Документ HTML опрацьовує браузер. Отже, усі інформаційні веб-ресурси мають однакову HTML-розмітку, що дає можливість відтворювати HTML-документ, за допомогою браузера, у звичному для користувача вигляді.

Наступним етапом є парсинг інформаційних веб-ресурсів – послідовний синтаксичний аналіз інформації, розміщеної на інтернет-сторінках. Текст інтернет-сторінок – ієрархічний набір даних, структурований за допомогою людських та комп'ютерних мов. Веб-парсинг використовується для автоматизованого збору контенту або даних із будь-якого інформаційного веб-ресурсу. Результат парсингу записується у базу даних, у файл зручного для читання чи обробки формату.

Для забезпечення коректності та актуальності інформації, отриманої різними пошуковими засобами, доцільно її зіставити та узагальнити за принципом "мажоритарного голосування". Основна його ідея полягає у порівнянні інформації з різних джерел і висуненні припущення, що ця інформація є коректною та актуальною. Принцип голосування «два з трьох» — будь-які два результати з трьох, що співпали, вважаються істиною. Можна використовувати співвідношення три з п'яти та інші.

Спираючись на вищенаведені результати, архітектуру програмної системи виявлення некоректної та неактуальної інформації із застосуванням принципу "мажоритарного голосування" представимо на рисунку 3.6.

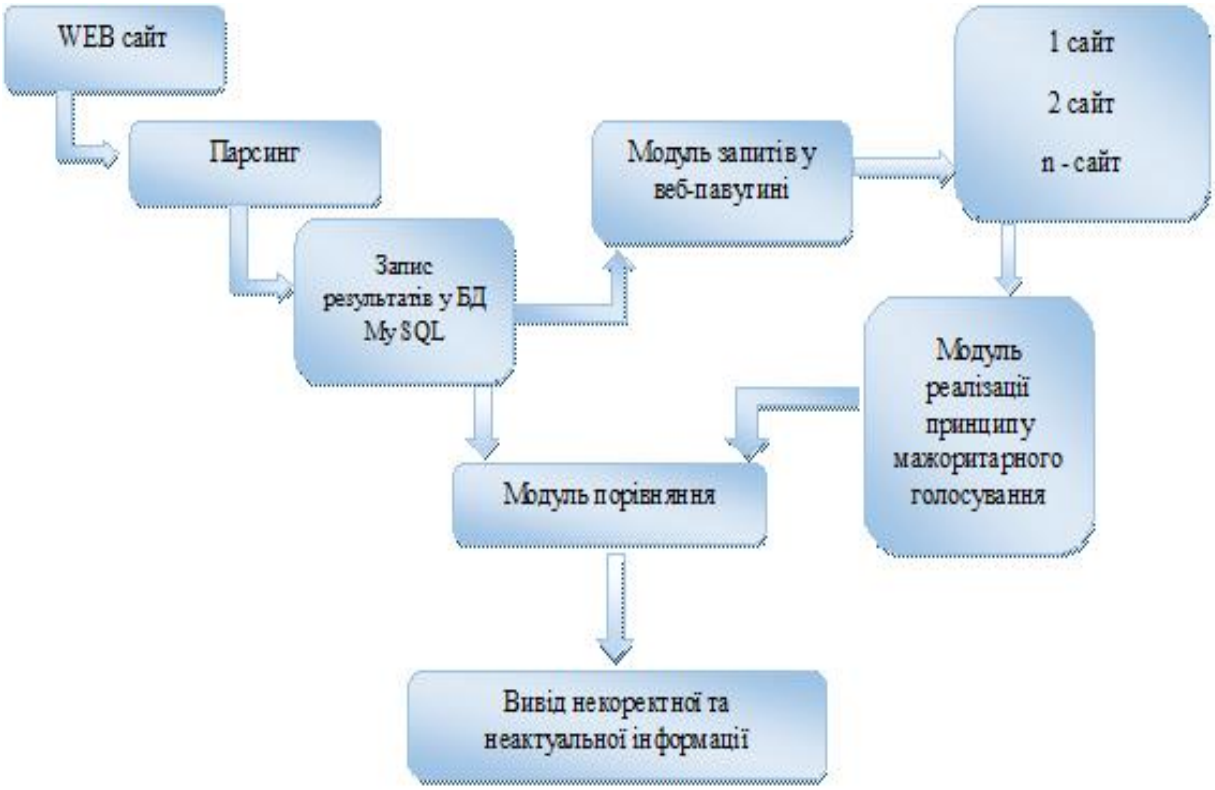


Рис. 3.6. Архітектура програмної системи виявлення некоректної та неактуальної інформації із застосуванням принципу “мажоритарного голосування”

На першому етапі проводиться парсинг інформаційних веб-ресурсів, запис результатів у базу даних, а також аналіз структури та заповнення цієї бази даних. На основі отриманих у ній записів, відбувається запит у веб-павутині, на що отримуємо контент із різнотипних інформаційних веб-ресурсів.

На другому етапі проводимо парсинг інформаційних веб-ресурсів, отриманих після запиту у веб-павутині, та записуємо у базу даних.

Структурувавши інформацію у базі даних, порівнюємо результати структурування за принципом “мажоритарного голосування”.

Отримавши результат, зіставляємо його із результатом, який отримали після парсингу інформаційних веб-ресурсів.

Для реалізації у роботі елементів штучного інтелекту та машинного навчання використано бібліотеку - PHP-ML.

Бібліотека підтримує версію інтерпретатора не менше 7.0 і додається до проекту через Composer.

Серед можливостей PHP-ML:

• робота з ML-алгоритмами;

• перехресна валідація;

• нейромережі;

• препроцессінг;

• видобування даних.

Для реалізація методу мажоритарного голосування в роботі реалізовано аналог функції для Python в PHP-ML:

```
use Phpml\Metric\ClassificationReport;

$actualLabels = ['name', 'titile', 'staff'];
$predictedLabels = ['name', 'staff', 'titile'];

$report = new ClassificationReport($actualLabels,
$predictedLabels);

$report->getPrecision();
// ['name' => 0.5, 'titile' => 0.0, 'staff' => 1.0]

$report->getRecall();
// ['name' => 1.0, 'titile' => 0.0, 'staff' => 0.67]

$report->getF1score();
// ['name' => 0.67, 'titile' => 0.0, 'staff' => 0.80]

$report->getSupport();
// ['name' => 1, 'titile' => 1, 'staff' => 3]

$report->getAverage();
// ['precision' => 0.5, 'recall' => 0.56, 'f1score' => 0.49]
```

Після створення звіту можна сформувати індивідуальні показники:

- precision (getPrecision()) – отримані екземпляри, які мають між собою звязки;
- recall (getRecall()) – відповідні екземпляри, які отримано.
- F1 score (getF1score()) – вимір точності в тестовій вибірці;
- support (getSupport()) – кількість перевірених елементів.

Отже, основними елементами алгоритмів пошуку контенту повинні бути: адаптація під структуру контенту типових інформаційних веб-ресурсів, виконання функцій, згаданих у вище розглянутих системах, концентрація на мовних особливостях, що використовуються для розмітки інформаційних веб-ресурсів, а також застосування моделей контенту, що ґрунтуються на онтологічному підході.

3.3. Метод виявлення некоректної та недостовірного контенту із використанням засобів машинного навчання

Структурування та розпізнавання текстової інформації є одним із напрямів інтелектуальних інформаційних систем. Основною компонентою таких систем є машинне навчання. Компоненти систем із засобами машинного навчання надають здатності програмним системам щодо автоматичного аналізу текстової інформації, з метою її структурування, виявлення некоректної та недостовірної інформації.

У дисертаційній роботі пропонується метод структурування та розпізнавання контенту веб-ресурсів, у який включено елементи машинного навчання. Такий підхід є основою процесу виявлення некоректної та застарілої інформації на веб-сайтах.

Розглянемо формальний опис системи, що виконує функції пошуку, структурування та розпізнавання текстової інформації на веб-ресурсах, записує у базу даних та порівнює її з еталонною. При цьому еталонна база даних може бути сформована на основі, знову ж таки, способу структурування контенту, знайденого на інших веб-ресурсах, та порівняння на принципах «мажоритарного голосування», розглянутих у попередньому підрозділі.

Для розв’язання цієї задачі необхідно використовувати методи, що базуються на основі комп'ютерної лінгвістики – напрямку штучного інтелекту, який ставить за мету використання математичних моделей для опису природних мов. Комп’ютерна лінгвістика частково перетинається із обробкою природних мов. Разом із тим, для застосування зазначених методів, спочатку необхідно запровадити формальне представлення контенту.

Отже, для виконання цього завдання, спочатку розглянемо спосіб представлення моделі контенту веб-сайту на основі онтологічного підходу, а потім проаналізуємо метод його формального опрацювання для виявлення некоректної та неактуальної інформації.

Під характеристиками конкретного об’єкта розуміємо деревоподібну систему значимих понять (концептів), які описують його певні значимі властивості. Сюди включаються як окремі типові характеристики, що використовуються для визначення його приналежності до певні групи, так і узагальнені характеристики для об’єкта в цілому. Тому об’єкти, що використовуються для аналізу достовірності їх наповнення, моделюємо наступною деревоподібною структурою

$$O_{Dsc} =< IdO, Pr\,O\,, Id\_Type, D_{Sc} >, \qquad (1)$$

яка включає ідентифікатори IdO об’єктів, ідентифікатор $Pr\,O$ батьківського елемента та атрибут Id_Type, що дозволяє відділити приналежність об’єкта до певної групи, яка вже володіє більш загальними характеристиками, ніж характеристика Dsc. При цьому загальний об’єкт відрізняється від часткових відсутністю батьківського, для якого $Pr\,O = NULL$.

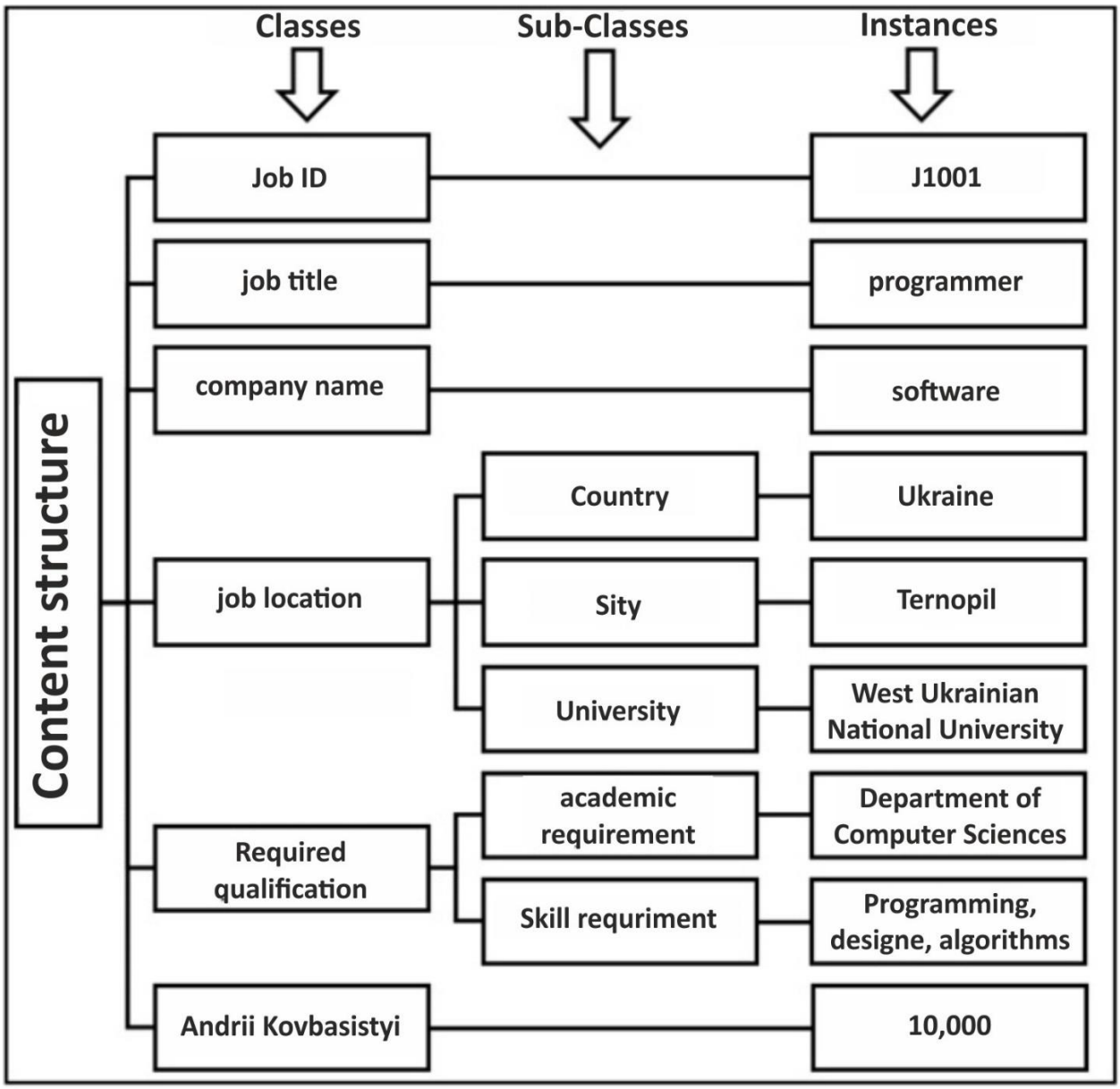


Рис. 3.7. Формалізоване подання аналізованої предметної області

На рисунку 3.7 представлено формалізоване подання аналізованої предметної області, а на рисунку 3.8 наведено графічне подання вищеописаної деревовидної структури об’єкта, що слугує основою для подальшого моделювання контенту.

Рис. 3.8. Графічне подання деревовидної структури контенту

Кожна характеристика аналізованого об’єкта допускає різні лінгвістичні представлення у вигляді словосполучень O_{Phr}. Слова, що визначають конкретні

характеристики, подаються у словоформах O_{Frm}, а також своїми основами O_{Bs}. Словоформи використовуються для представлення понять користувачам, а основи – для автоматичної ідентифікації еквівалентності мовних представлень між аналізованими та еталонними даними. Атрибути введених понять згрупуємо у наступні структури:

$$O_{Bs} =< IdLg, IdBs, WBase >, \quad (2)$$

$$O_{Frm} =< IdLg, IdFm, IdBs, WForm >, \quad (3)$$

$$O_{Phr} =< IdCn, IdLg, IdPh, IdBs, IdFm, Id\ Pr\ B\ s >, \quad (4)$$

де $IdLg$ – ідентифікатор мови реалізації, $IdBs$ – ідентифікатор основи слова, $WBase$ – основа слова, $IdFm$ – ідентифікатор форми слова, $WForm$ – форма слова, IdO – ідентифікатор об'єкта, $IdPh$ – ідентифікатор характеристик, $Id\ Pr\ B\ s$ – ідентифікатор батьківської основи об'єкта.

Для порівняння значень аналізованих характеристик об'єктів та їх відповідних еталонних даних, необхідно здійснювати аналіз інформації, представленої на відповідних веб-сторінках.

Структура опису інформації про аналізовані об'єкти є наперед невідомою і може бути різною, залежно від тематики, наповнення, а також технічної реалізації цих сайтів.

Доцільно, щоб інформація на веб-сторінках була структурована, а не просто розбита на параграфи чи абзаци. Така вимога дозволяє значно спростити процес аналізу встановлення коректності та актуальності контенту. У даному випадку під структурованістю мається на увазі оформлення інформації у вигляді визначених структур.

Для аналізу інформації, що стосується конкретного об'єкта, формуємо множину ключових характеристик $KDrO$. Для підтримки аналізу вмісту веб-сторінок створено наступну допоміжну структуру AS:

$$AS =< IdPg, IdO, IdIt, IdBs, IdFm, Id\ Pr\ B\ s > \quad (5)$$

де $IdPg$ – ідентифікатор аналізованої веб-сторінки, IdO – ідентифікатор аналізованого об'єктА , $IdIt$ – ідентифікатор характеристики.

При аналізі $HTML$ -коду чергової веб-сторінки спеціалізованого веб-сайту встановлюємо її ідентифікатор:

$$CurPgId := max\left(\pi_{IdLPg}(BF)\right) + 1 \quad (6)$$

Далі виділяємо елементи $LSTIt$ відповідних характеристик, що наповнюють визначені теги. Ці елементи в подальшому використовуються для порівняння із елементами еталонних характеристик об'єктів OE_{Dsc} конкретної бази даних. OE_{Dsc} описується наступною структурою

$$OE_{\text{Dsc}} =< IdO, Pr\,O\ ,\text{Id\_}TypeE, D_{\text{SCE}} > \quad (7),$$

де Id_$TypeE$ – тип еталонних характеристик, D_{SCE} – еталонні характеристики.

Нехай $Wrd_k(It_{Pg,Lst})$ – деяке k та характеристика, виділена із відповідного об'єкта $It_{Pg,Lst}$. Якщо вона співпадає із коректною та актуальною інформацією, то її Dsc співпадає із відповідними коректними та актуальними значеннями:

$$BsWrd = \pi_{\text{Dsc}}\left(\sigma_{DscE=Dsc(Wrd_k(It_{Pg,Lst}))}(O_{Dsc}, OE_{Dsc})\right) \quad (8)$$

Якщо значення цих відповідних характеристик не співпадають, то дана інформація потребує оновлення.

Отже, представлений математичний опис контенту та формальний метод його опрацювання дає можливість створити програмну систему для розпізнавання контенту, його структурування і виявлення у ньому некоректної та неактуальної інформації.

3.4. Архітектура програмної системи для реалізації методу виявлення некоректного та недостовірного контенту

Спираючись на аналіз існуючих систем контент-аналізу веб-сайтів, можемо сформулювати основні вимоги до алгоритмів автоматизованого виявлення застарілої та недостовірної інформації.

Серед переліку вимог:

- налаштування фільтрів URL, щоб не парсити зайві сторінки;
- налаштування глибини парсингу;
- якісне скачування контенту;
- обробка у декілька потоків та збереження контенту у різних форматах.

Після здійснення парсингу згенеровані результати необхідно опрацювати, щоб надати інформації вигляду, придатного для подальшого використання. Конкретний формат залежить від того, як у подальшому будуть оброблятися зібрані дані. Доволі часто з парсенного контенту за допомогою XML формується RSS-потік, який зручний для використання даних без процедури рерайтингу. Іноді результат парсингу розміщують у CSV-файл, оскільки цей текстовий формат дуже простий у подальшій обробці, легко конвертується у SQL-запити і без проблем відкривається в Excel. В особливих випадках потрібно, щоб кінцеві дані були представлені у вигляді електронних таблиць XLS.

Розглянемо приклад застосування розробленої системи для структурування, розпізнавання, виявлення некоректної, неточної та недостовірної інформації про кадрове забезпечення на веб-ресурсі навчального закладу. За основу прикладу візьмемо сайт факультету комп’ютерних та інформаційних технологій ТНЕУ.

Приклад реалізації узагальненого алгоритму структурування контенту за допомогою парсингу наведено на рис. 3.9.

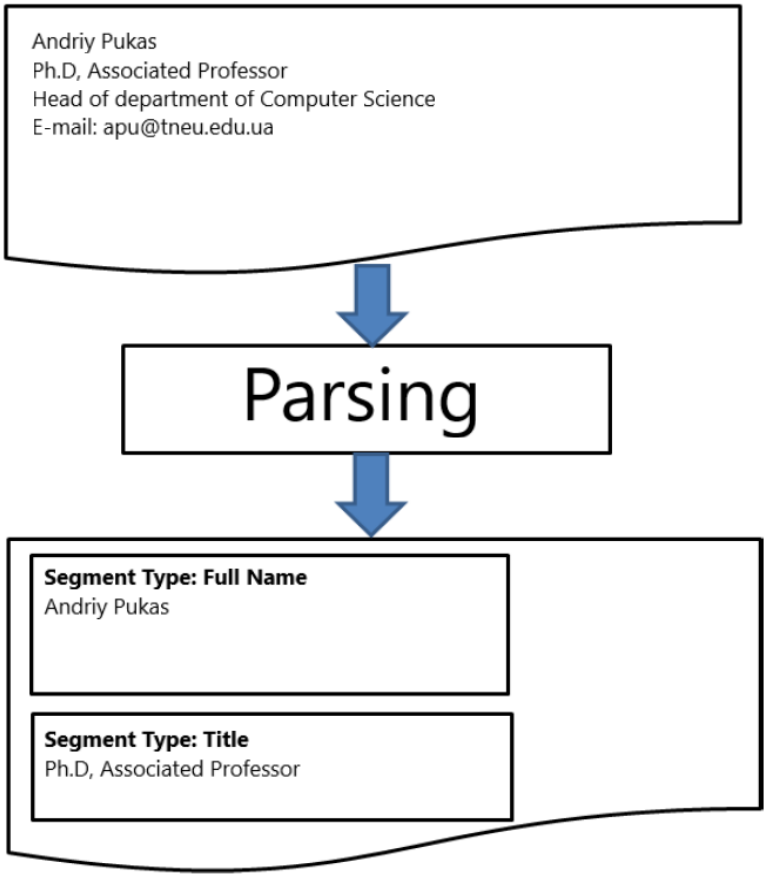


Рис. 3.9. Приклад реалізації узагальненого алгоритму структурування контенту за допомогою парсингу

Як бачимо, текстовий контент «Andriy Pukas, Ph.D, Associated Professor» структуровано на два сегменти: «Andriy Pukas», «Ph.D, Associated Professor».

На рис. 3.10 наведено приклад виконання узагальненого алгоритму реалізації методу виявлення неактуальної та недостовірної інформації на веб-ресурсах.

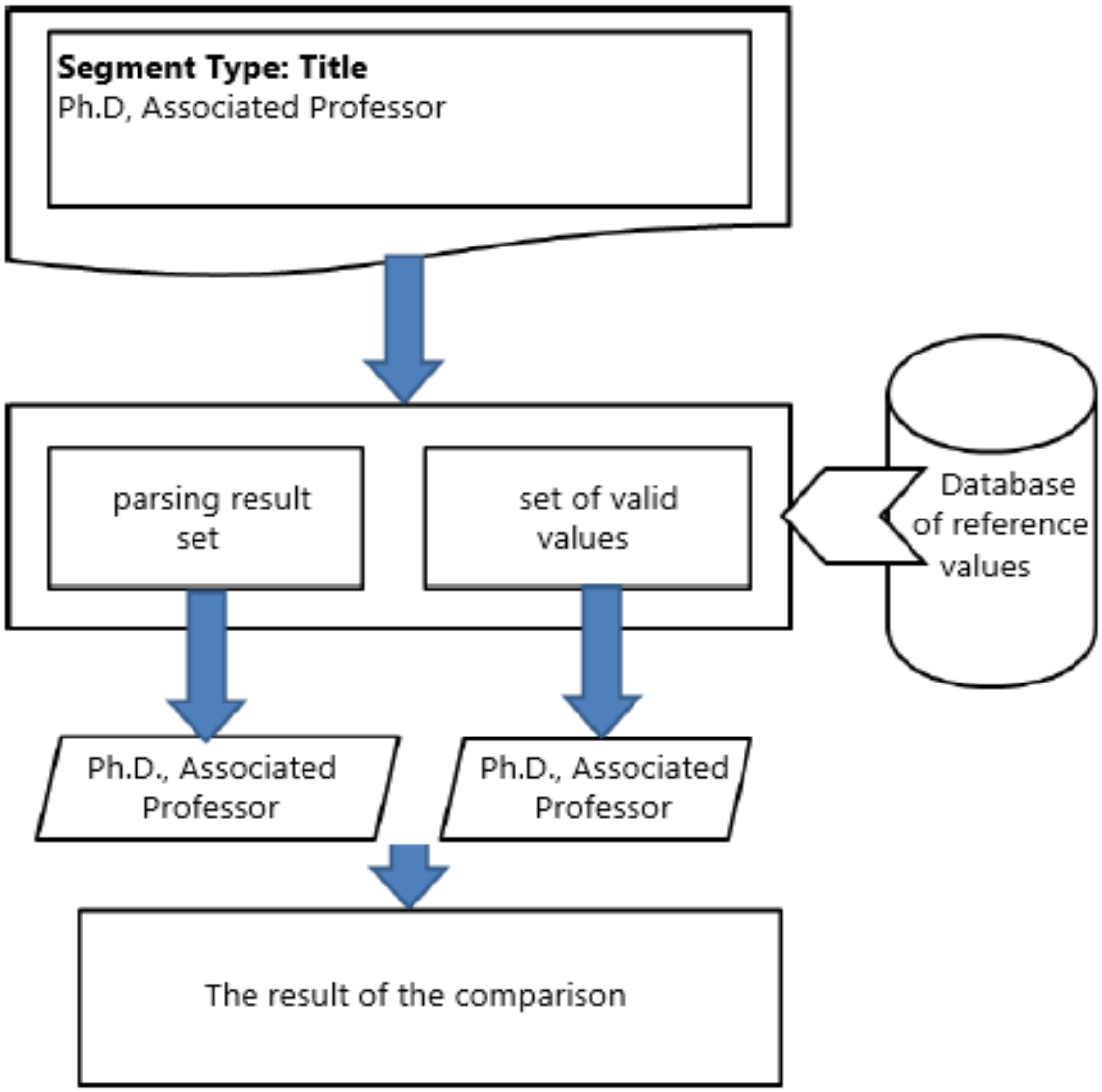


Рис. 3.10. Узагальнений алгоритм реалізації методу виявлення неактуальної та недостовірної інформації на веб-ресурсах і порівняння із існуючою базою даних

У процесі аналізу текстової інформації, що отримана з веб-сторінок, необхідно приділити увагу процесу нормалізації тексту. Це дозволить підвищити відсоток встановлення еквівалентної інформації між аналізованими та еталонними даними. У нормалізації тексту деякі з назв сутностей перетворюються, щоб зробити їх послідовними. Розглянемо суть цього процесу на прикладах.

У Таблиці 3.1 презентовано перетворення, що виконуються на декількох текстових фразах.

На етапі нормалізації розширюємо частину скорочень, використовуючи відповідні предметні довідники. У Таблиці 3.2, наприклад, абревіатура «кан.тех. наук» розширена як «кандидат технічних наук». Ми також трансформуємо частину тексту в нову форму. Приміром, ініціали людини, що ідентифікуються в процесі парсингу, перетворюються наступним чином, так, як показано у Таблиці 3.1.

Деякі фрази також трансформуються на більш звичні відповідники, що використовуються в тій чи іншій предметній області. Наприклад, «керівник кафедри» перетворюємо на «завідувач кафедри». Також приклад трансформації фраз на більш звичні відповідники, що використовуються у тій чи іншій предметній області, у вигляді скорочень, показано в Таблиці 3.2.

Таблиця 3.1.

Приклад нормалізації текстової інформації

Input	Output	Type
кан.тех. наук	кандидат технічних наук	Degree
ANDRIY PUKAS	Andriy Pukas	Name
керівник кафедри	завідувач кафедри	Staff

Таблиця 3.2.

Приклади перетворення фраз на більш звичні відповідники у процесі нормалізації текстової інформації

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

Висновки до третього розділу

1. Сформульовано основні вимоги до засобів автоматичного аналізу контенту, основним завданням яких є виявлення застарілої та некоректної інформації. Встановлено, що засоби повинні бути адаптовані на типові структури веб-сторінок, а також на відомі структури опису веб-сторінки мовою розмітки HTML. Також показано основні складнощі щодо виявлення тегів із різними іменами та порівняння результатів із відомими даними, записаними в базах даних організацій чи виявлені на інших веб-ресурсах

2. Запропоновано та обґрунтовано підхід до побудови контент аналізу з автоматичним виконанням функцій. Основна ідея побудови системи полягає у перетворенні інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, що існують в організаціях або виявлені на інших веб-ресурсах, з їх вибором за принципами «мажоритарного голосування» та з подальшим порівнянням. Наведено приклади використання запропонованого підходу автоматизованого пошуку застарілої та некоректної інформації на сайтах закладів вищої освіти, де було отримано доступ до відповідних баз даних.

3. Запропоновано та обґрунтовано новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання. За рахунок уведення формального опису контенту та компонентів машинного навчання, рекомендований метод, порівняно із вже відомими, забезпечує автоматизоване структурування контенту, виявлення у структурі контенту недостовірної, неточної чи неактуальної інформації.

РОЗДІЛ 4.
ПРИКЛАДНА ПРОГРАМНА СИСТЕМА ДЛЯ ПІДТРИМКИ ФУНКЦІОНУВАННЯ ІНФОРМАЦІЙНИХ ВЕБ-РЕСУРСІВ

У попередніх розділах отримано наукові результати, що стосуються нових методів опрацювання контенту. Для реалізації цих методів створено прикладну програмну систему для підтримки інформаційних веб-ресурсів. Спочатку сформульовано вимоги до прикладної системи, обґрунтовано її архітектуру та спроектовано усі необхідні модулі. Наведено результати проектування інтерфейсу користувача. Потім подано результати використання розробленої прикладної програмної системи для розв’язування ряду прикладних задач, таких як: підтримка функціонування інформаційного веб-ресурсу Західноукраїнського національного університету; дослідження та моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів на прикладі Центру надання адміністративних послуг Тернопільської міської ради та ТОВ «Терногаз». У завершальній частині розділу для кожного з прикладів проведено дослідження ефективності розробленої системи, відповідно до розробленої моделі ефективності.

Результати досліджень цього розділу опубліковано в працях[5-7].

4.1. Функціональні можливості

За результатами теоретичних досліджень у нашій дисертаційній роботі розв’язано ряд прикладних задач, пов’язаних зі структуруванням контенту, виявленням недостовірної та неактуальної інформації на сайтах-візитівках. Основним інструментом для вирішення цих завдань є створена прикладна програмна система для підтримки функціонування інформаційних веб-ресурсів.

Узагальнену архітектуру програмної системи структурування та розпізнавання текстового контенту веб-ресурсів наведено на рис. 4.1.

Основні бізнес-процеси у рамках проектованої системи для структурування та розпізнавання текстового контенту веб-ресурсів можна описати наступною послідовністю кроків:

1. Користувач створює запит, який надсилається до аналізатора запитів.

2. У блоці аналізатор запитів розділяється на певні категорій та розбивається за ключовими словами.

3. У наступному етапі, використавши опис ключа запиту, створюються запити у декілька пошукових системах та визначений інформаційний веб-ресурс.

4. Збираються та порівнюються отримані дані. Результати записуються у базу даних для порівняння із еталонними.

5. Здійснюється перевірка отриманих та еталонних результатів. У випадку виявлення розбіжностей виводиться повідомлення. Якщо такий запис відсутній у таблиці з еталонними даними, він додається.

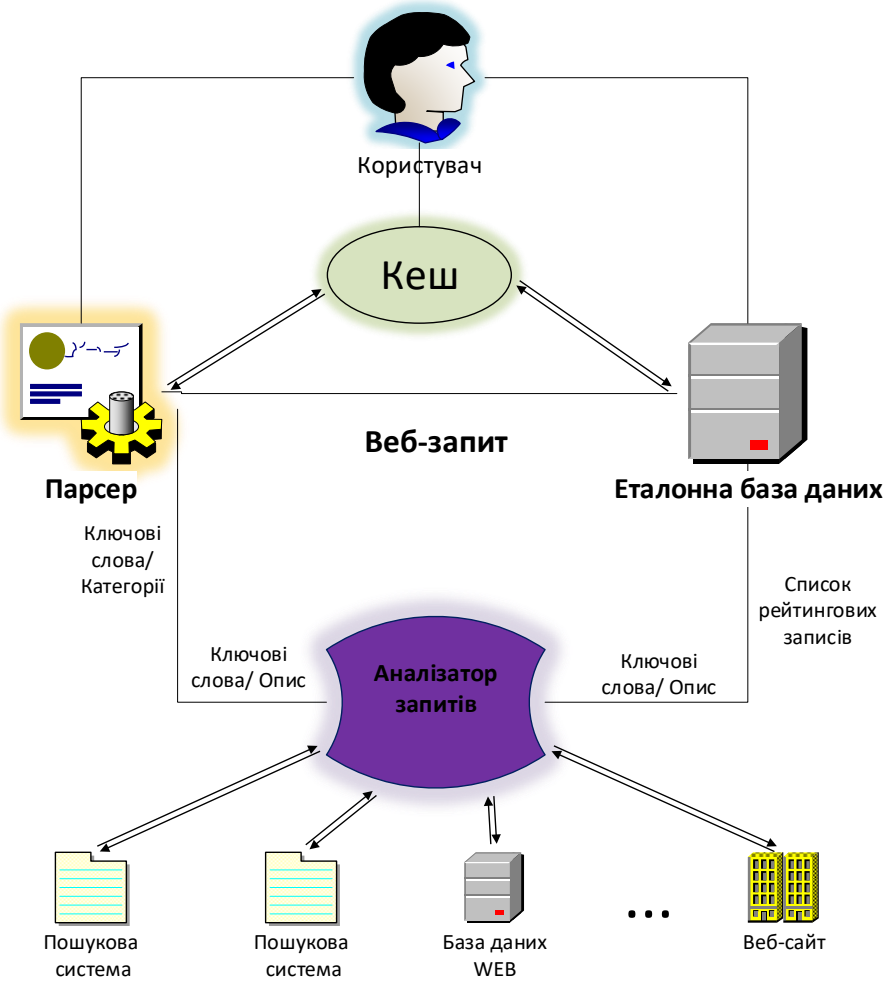


Рис. 4.1. Архітектура для реалізації програмної системи структурування і розпізнавання текстового контенту веб-ресурсів.

На рисунку 4.2 представлено Use-case діаграму системи структурування і розпізнавання застарілої та неактуальної інформації.

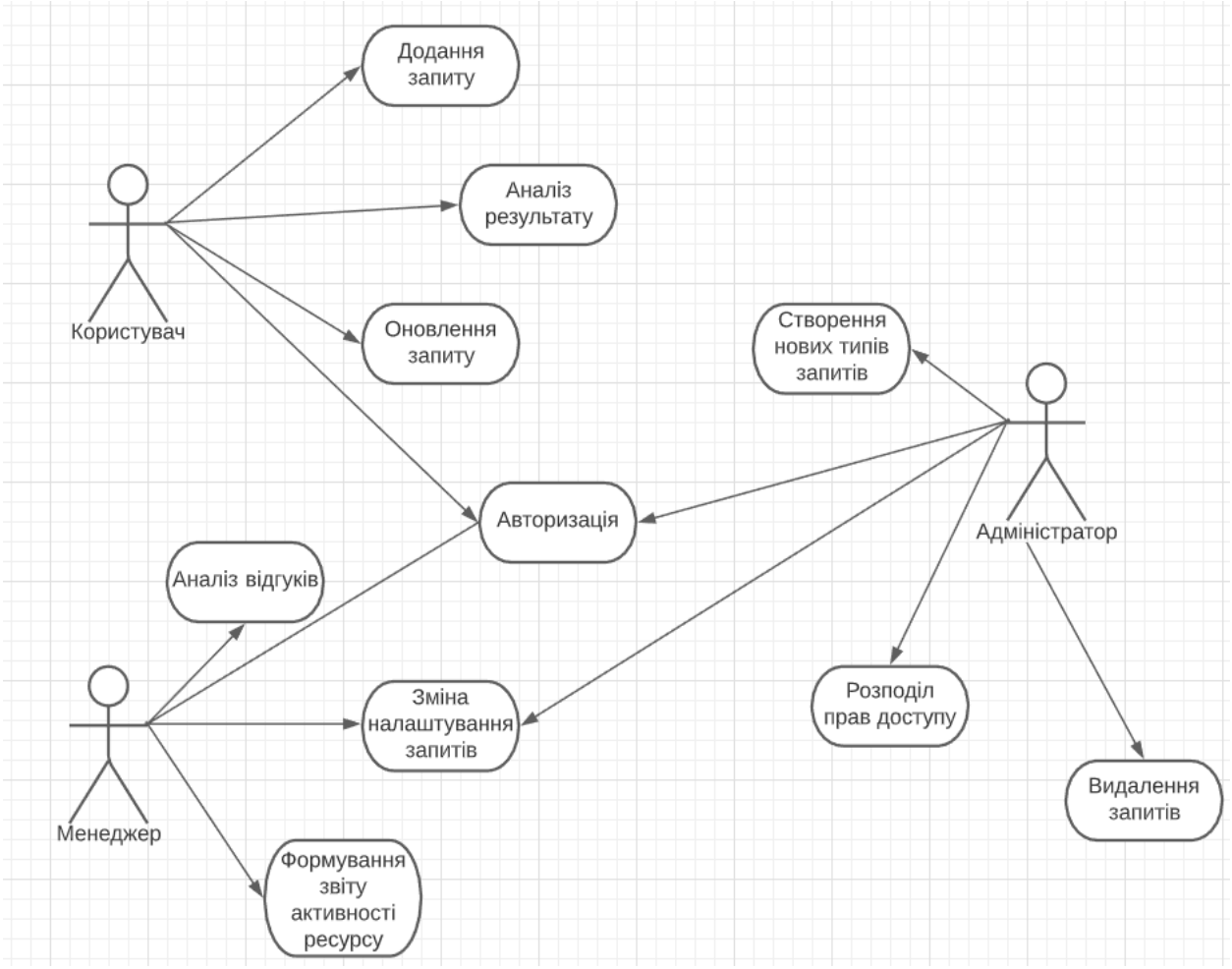


Рис. 4.2. Діаграма варіантів використання створеної програмної системи

У процесі виявлення акторів та сценаріїв використання, що застосовуються на етапі проектування системи, виділено наступні групи:

1. Адміністратор – користувач, який повністю здійснює управління системою. У таблиці 4.1 представлено опис варіантів використання для даного користувача.

2. Менеджер – користувач, який може змінювати налаштування запитів та формує звіт активності ресурсів. У таблиці 4.2 представлено опис варіантів використання для даного користувача.

3. Користувач, який може додавати та аналізувати запити. У таблиці 4.3 представлено опис варіантів використання для даного користувача.

Таблиця 4.1

Опис варіантів використання для адміністратора

Актори	Можливі варіанти використання
Адміністратор	Авторизація
	Створення нових типів запитів
	Зміна налаштування запитів
	Розподіл прав доступу
	Видалення запитів

Таблиця 4.2

Опис варіантів використання для менеджера

Актори	Можливі варіанти використання
Менеджер	Аналіз відгуків
	Авторизація
	Зміна налаштування запиту
	Формування звіту активності ресурсів

Таблиця 4.3

Опис варіантів використання для користувача

Актори	Можливі варіанти використання
Користувач	Додання запиту
	Аналіз результатів
	Оновлення запитів
	Авторизація

Приклади використання аналізуються, щоб отримати короткий опис того, як застосовуються варіанти використання для запропонованої системи. Опис використання містить додаткову інформацію, що дозволяє користувачам простіше зрозуміти предметну область.

Такі описи в основному складаються з трьох основних частин: оглядова інформація, взаємозв’язки та потік подій. У таблицях 4.4 – 4.6 представлено основні описи всіх варіантів використання.

Таблиця 4.4.

Use-case опис-запис відвідуваності

Use-case назва	Записувати відвідуваність за допомогою тегів
Головний актор	Користувач
Рівень важливості	Високий
Короткий опис	Користувач авторизується
Потік подій	1. Користувач реєструється. 2. Адміністратор підтверджує реєстрацію. 3. Менеджер надає доступ відповідно до прав доступу. 4. Користувач здійснює аналіз контенту та його співставлення.

Модель використання та її описи є основними елементами, за допомогою яких розробляється та будується система.

Таблиця 4.5.

Use-case опис – звіт авторизації користувачів

Use-case назва	Звіт про авторизацію користувачів
Головний актор	Менеджер
Рівень важливості	Високий
Короткий опис	Створення звіту використання ресурсу та його ефективності. Звіт містить опис сесії, форми запитів та їх успішність
Потік подій	1. Формує звіт активності на ресурсі. 2. Зміна налаштування запиту.

Таблиця 4.6.

Use-case опис – керування правами доступу

Use-case назва	Керування правами доступу
Головний актор	Адмінстратор
Рівень важливості	Високий
Короткий опис	Адміністратор керує правами доступу користувачів, які зареєстровані. Створення нових форм запитів, редагування існуючих та видалення застарілих.
Потік подій	1. Авторизація користувачів. 2. Адміністратор надає права доступу для визначених користувачів. 3. Створення нових типів запитів. 4. Налаштування запитів. 5. Видалення запитів.

На рисунку 4.3 представлено структуру програмних модулів інтелектуальної системи аналізу структурованості веб-ресурсів. Серед наведених модулів ключовими є manager, model.

Рис. 4.3. Структура модулів програмної системи

На рис. 4.4 наведено діаграму компонентів програмної системи, серед яких потрібно виділити наступні:

1. User GUI – компонент, що відображає інтерфейс роботи звичайного користувача із системою. Цей компонент реалізований з використанням HTML, CSS, Bootstrap, Javascript, jQuery.

2. Administrator GUI – компонент, що відображає інтерфейс роботи адміністратора із системою. Цей компонент реалізований з використанням класичних засобів для розробки веб-інтерфейсів. Логіка взаємодії із серверною підсистемою реалізована із використанням PHP та фреймворку Laravel.

3. Компонента, яка використовується для зберігання інформації, реалізована засобами СУБД MySQL.

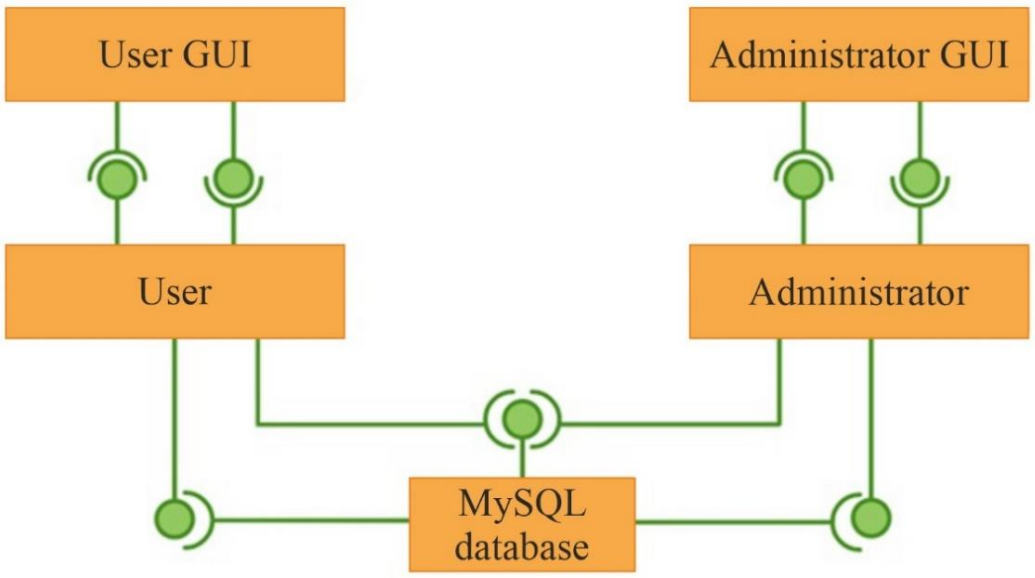


Рис. 4.4. Діаграма компонентів програмної системи

Як бачимо, результатом функціонування програмної системи, архітектуру якої подано на рис. 4.1, є структурована інформація. Результати структурування представлено схемою організації бази даних на рисунку 4.5.

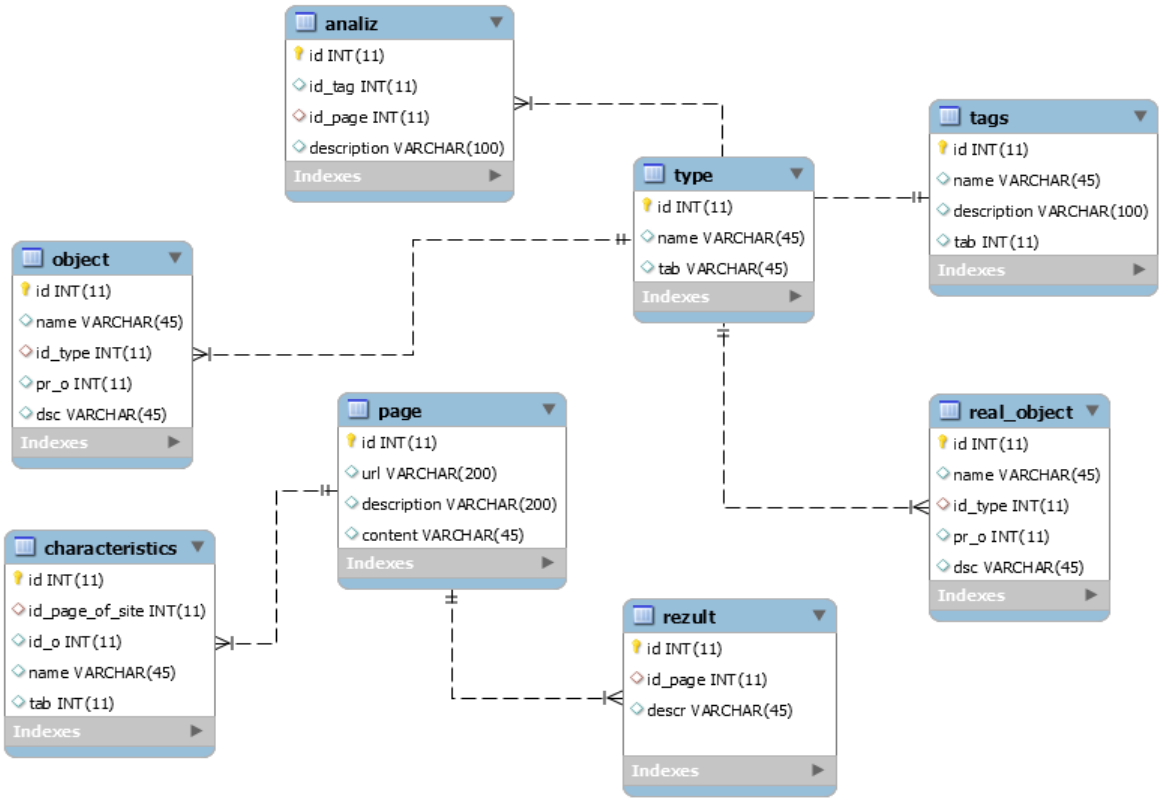


Рис. 4.5. ER-діаграма бази даних обробки та зберігання інформації для програми виявлення неактуальної та недостовірної інформації на веб-ресурсах і порівняння із існуючою базою даних

На рис. 4.6-4.10 наведено структури кожної сутності спроектованої бази даних.

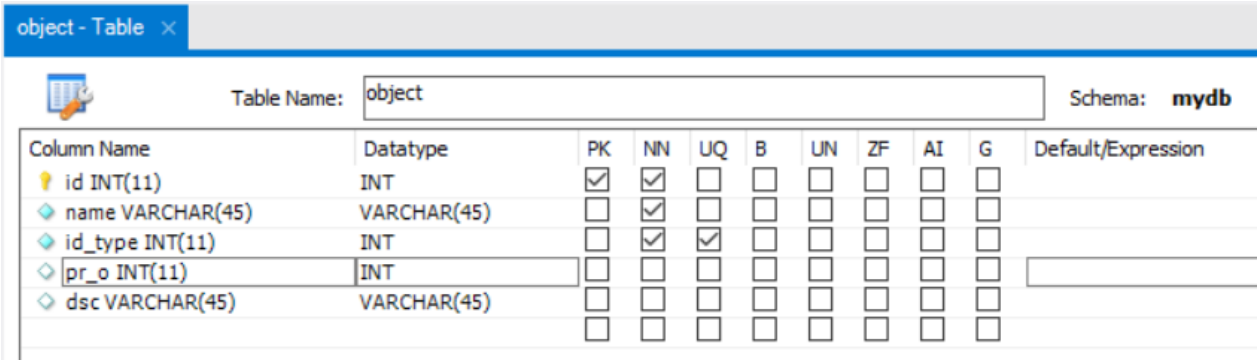


Рис. 4.6. Опис відношення об’єкта, стосовно якого здійснюється запит для пошуку

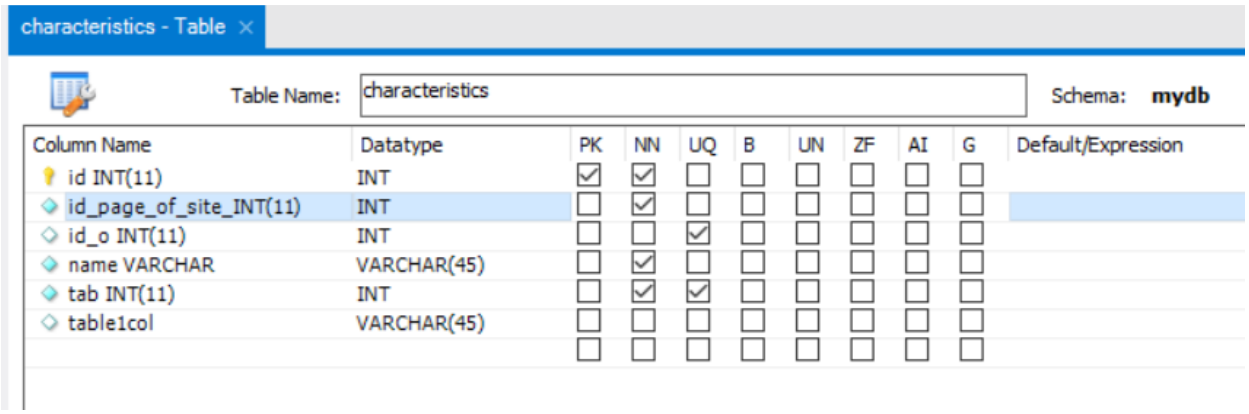


Рис. 4.7. Структура бази даних властивостей інформаційного веб-ресурсу

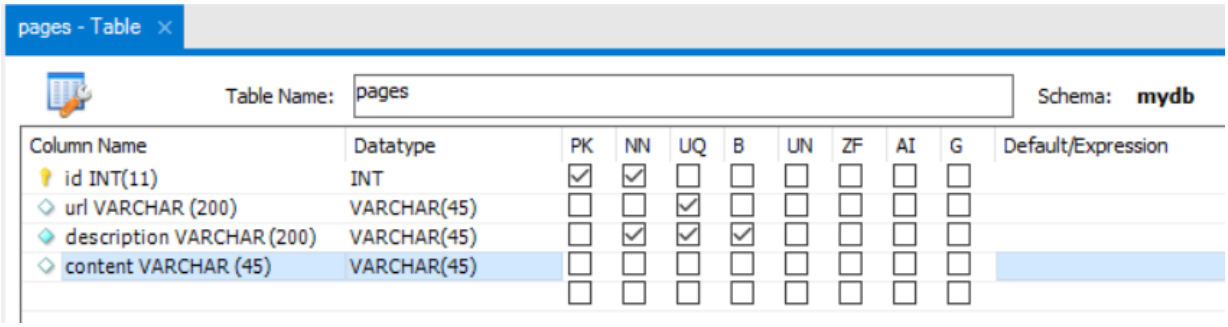


Рис. 4.8. Структура бази даних сторінки інформаційного веб-ресурсу

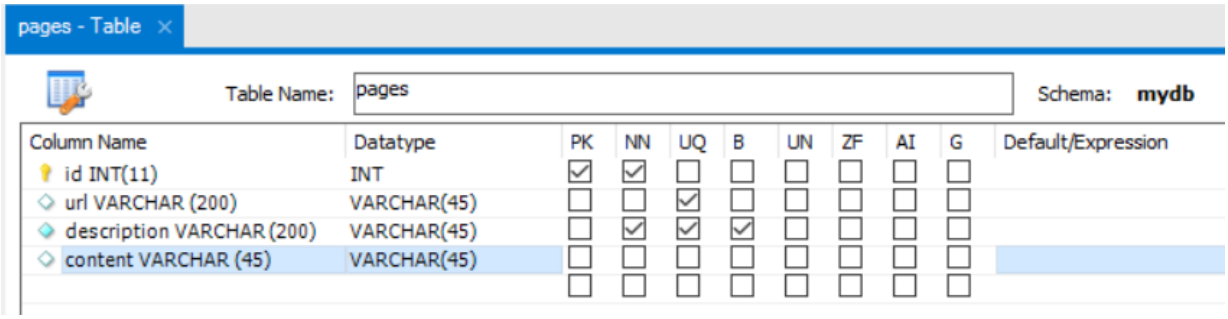


Рис. 4.9. Структура отриманих даних запиту до інформаційного веб-ресурсу

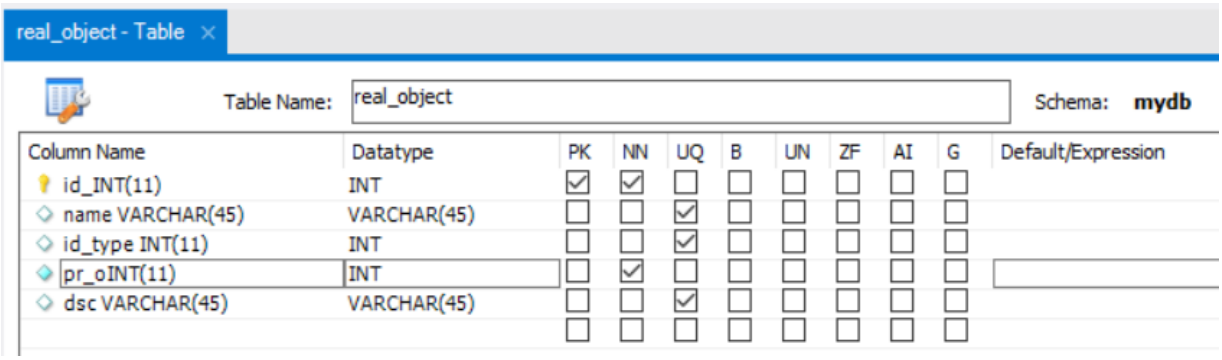


Рис. 4.10. Структура бази даних із еталонними даними

Також у структурі бази даних є інші таблиці, серед яких варто відзначити analize, tags, rezult.

Відношення «analize» використовується для зберігання інформації про основні структурні елементи, що необхідні для здійснення процедури аналізу контенту. Відношення «tags» застосовуються для опису структури HTML-тегів та інформації яка зберігається між ними. Відношення «rezult» використовується для представлення результатів аналізу.

Програмна система, запропонована у даній роботі, володіє кількома перевагами: гнучкістю, інтегрованістю та швидкодією за рахунок використання сучасних програмних засобів.

4.2. Особливості організації інтерфейсу та функціонування

Робота із системою розпочинається із завантаження початкової сторінки яка завантажується за вказаною адресою. На рисунку 4.11 представлено початкову сторінку системи.

Серед основних елементів управління, які подані на даній сторінці, необхідно відзначити:

1. Search-стрічка для вказування адреси аналізованого веб-ресурсу. Як адресу можна використовувати увесь ресурс або окрему його веб-сторінку.

2. Для управління процесом відображення результатів пошуку використовуються параметри sample enabled для простого пошуку, visuals enabled для візуалізації у вигляді схем та діаграм.

3. Для збереження результатів пошуку задіюються інструменти, що

дозволяють їх опрацьовувати у вигляді csv.excel, json.

4. У процесі опрацювання аналізованих запитів важливою характеристикою є час аналізу та час відкликання інформаційного веб-ресурсу. Для зручності такого аналізу реалізовано окремий програмний інструментарій ER Wait Times.

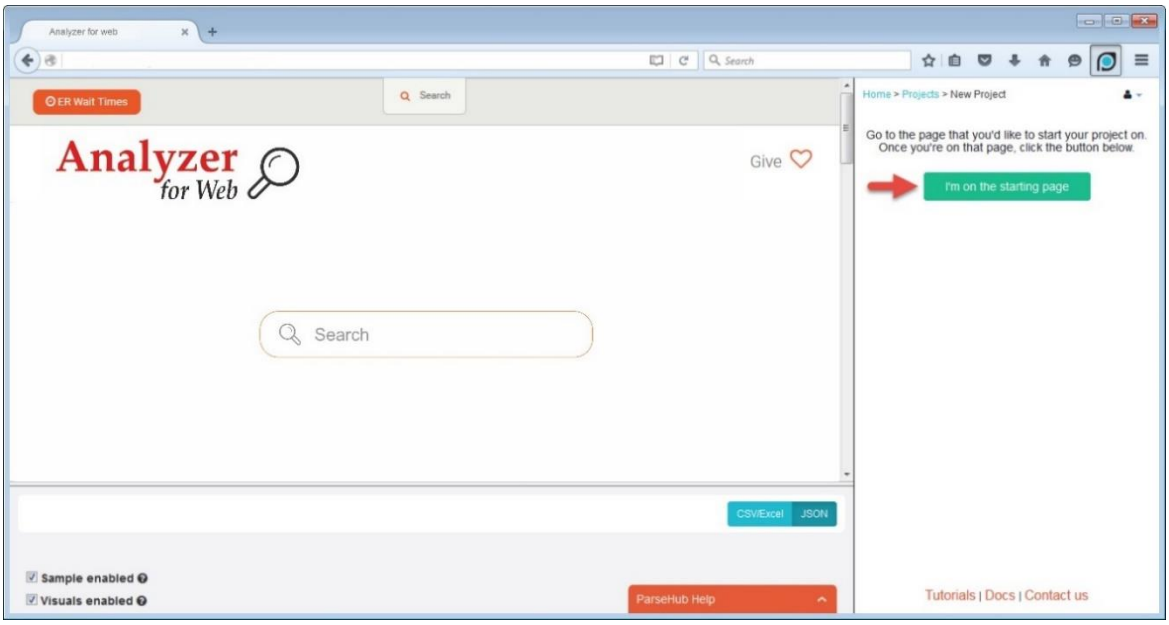


Рис. 4.11. Початкова сторінка системи

Для спрощення роботи користувача із системою реалізовано довідковий модуль Parser Help. Його структура побудована за принципом wiki help, що досить зручно для формування бази знань із описами основних принципів роботи із системою.

Наступним етапом роботи із системою є вибір елементів або ж структури контенту для аналізу та порівняння із еталонними даними. На рисунку 4.12 представлено сторінку для вибору елементів для аналізу та їх порівняння.

Серед основних елементів управління, які представлені на даній сторінці, необхідно відзначити:

1. Інструмент для перевірки правильності та достовірності інформації (перевірка за типами даних, розмірність полів, розмір графічних об’єктів, перевірка імен файлів на коректність).

2. Information is valid – вибір об’єкта для дослідження (визначається

структура аналізованих об’єктів). Його величину визначає сам користувач.

3. Select attributes – застосовується для вибору атрибутів досліджуваного об’єкта. Використовується для уточнення та вибору певних атрибутів, що цікавлять користувача.

4. View Profile – застосовується для валідації даних із можливістю оновлення її у базі даних.

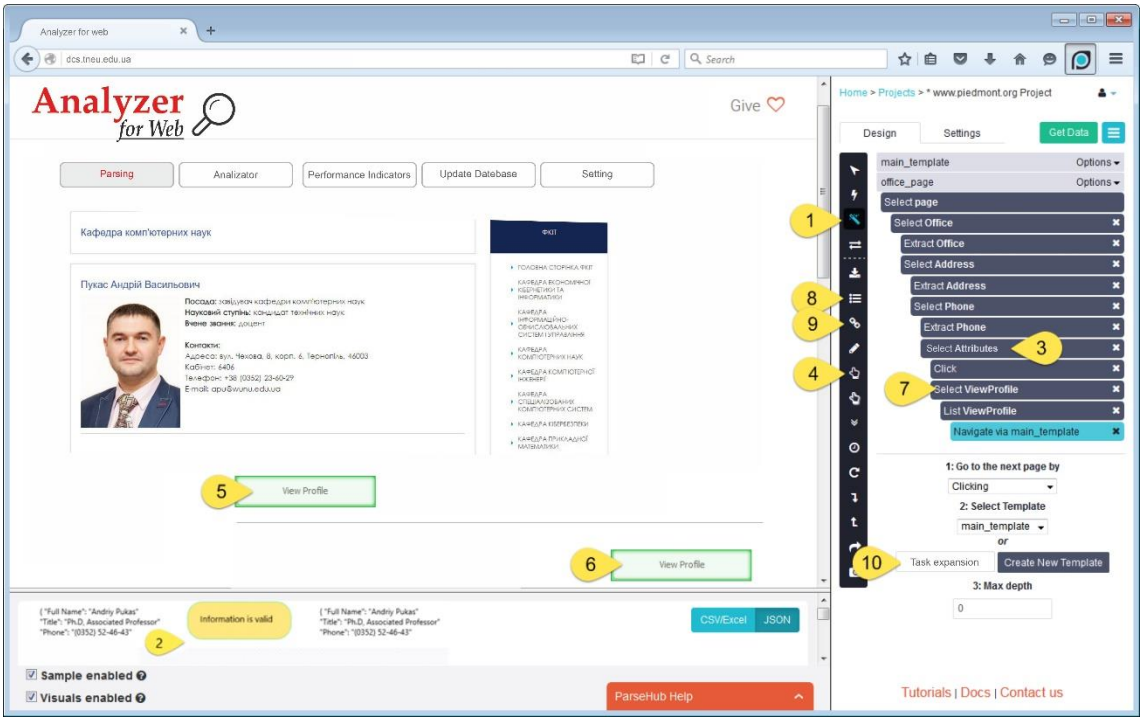


Рис. 4.12. Сторінка із представленням результатів парсингу та описом основних інструментів їх аналізу

У випадку браку результатів виконання, практикується доповнення правил (парсинг, підключення словників) та меж аналізованої області за допомогою інструменту task expansion.

Наступним етапом роботи системи - оновлення інформації у базі даних, якщо виявлена некоректна або застаріла інформація. На рисунку 4.13 представлено сторінку для вибору параметрів оновлення бази даних.

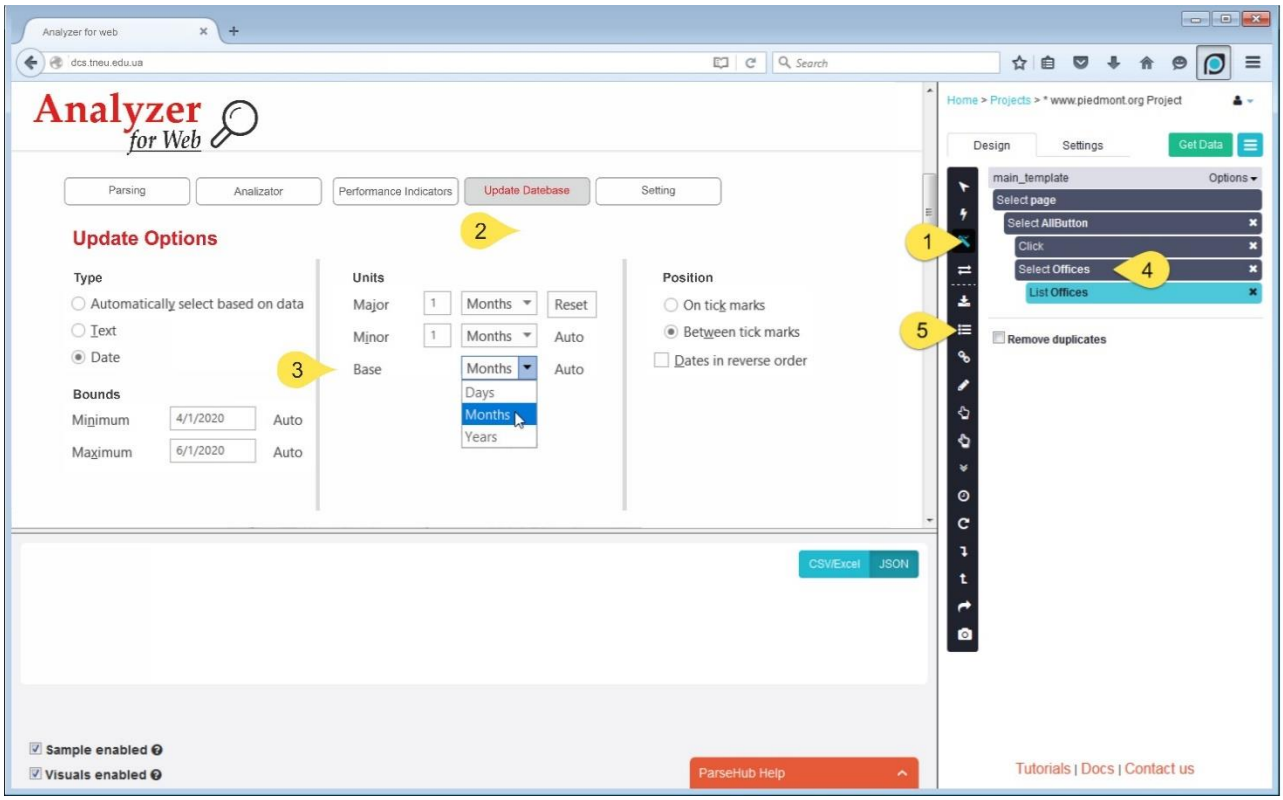


Рис. 4.13. Сторінка оновлення бази даних на основі результатів аналізу

Серед основних елементів управління, які представлені на даній сторінці, необхідно відзначити:

1. Type – використовується для вибору типу даних для оновлення. Text, date – для вибору типу даних для оновлення.

2. Для задання інтервалу часу для оновлення у базі даних застосовується властивість bounds. Bounds використовується для зменшення часу пошуку системою у базі даних.

3. Властивість Units використовується для уточнення важливої дати (Major) та другорядної дати (Minor) для оновлення.

Наступним етапом роботи із системою є оцінка ефективності роботи веб-ресурсу. На рисунку 4.14 представлено сторінку щодо вибору параметрів для аналізу ефективності роботи веб ресурсу.

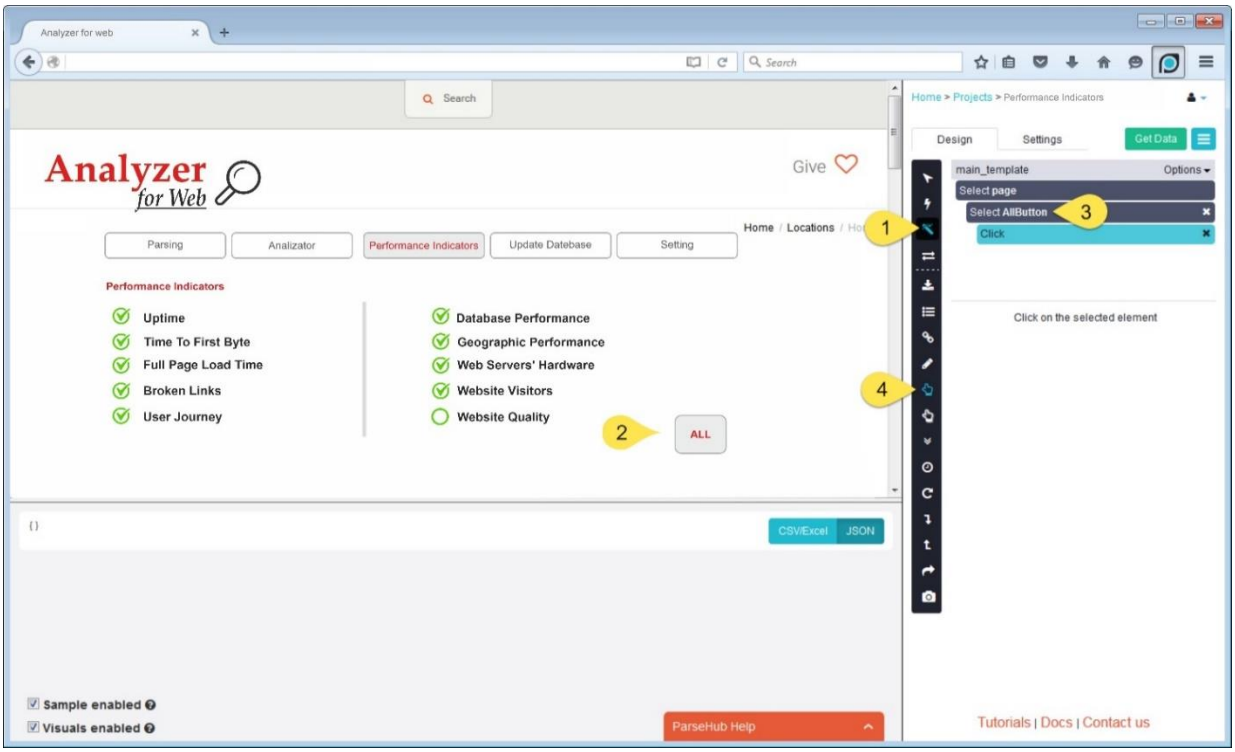


Рис. 4.14. Сторінка для оцінка ефективності роботи веб-ресурсу, виходячи із результатів структурованості контенту

Під час оцінки ефективності роботи веб-ресурсів, виходячи із результатів структурованості контенту, виділено наступні показники:

1. Uptime – параметр безперебійної роботи веб ресурсу.
2. Показник для визначення часу завантаження стартової сторінки, що відображає тривалість очікування користувачем, Time To First Byte.
3. Full Page Load Time – час завантаження вихідного коду обраної сторінки.
4. Broken Links – недіючі посилання, що впливає на якість просування веб-ресурсу у рейтингах.
5. Database Performance – застосовується для відстежування продуктивності бази даних, щоб забезпечити безперебійну роботу веб-сайту. Багато веб-сайтів містять динамічний вміст, вилучений із бази даних, що є причиною повільної реакції веб-сайту та погана робота бази даних.
6. Geographic Performance – швидкість та доступність веб-сайту з різних куточків світу (за допомогою PRTG Cloud Ping Sensor або Cloud HTTP

Sensor).

7. Web Servers' Hardware – файли журналів, записи бази даних, фотографії та відеофайли займають значну кількість дискового простору веб-серверів, які значно уповільнюють завантаження веб-ресурсу.

8. Для визначення показник ефективності веб-ресурсу, а саме відвідуваність застосовується Website Visitors.

Наступним етапом роботи із системою є можливість налаштування певних параметрів для видобування інформації із веб-ресурсів. На рисунку 4.15 представлено сторінку для вибору параметрів оновлення бази даних. Серед основних елементів управління, які подано на даній сторінці, необхідно відзначити:

1. Number of threads – чим більша кількість потоків, тим більше URL-адрес можна опрацювати за одиницю часу. Однак більша кількість потоків призводить до більшої кількості використовуваних ресурсів сервера. Рекомендованою кількістю потоків є 10–15.

2. Scan Time – встановлюється обмеження часу для сканування веб ресурсу. Одиниця виміру – година.

3. Maximum depth – цей параметр використовується для вказівки глибини сканування веб-ресурсу. Початкова сторінка має рівень вкладеності 0.

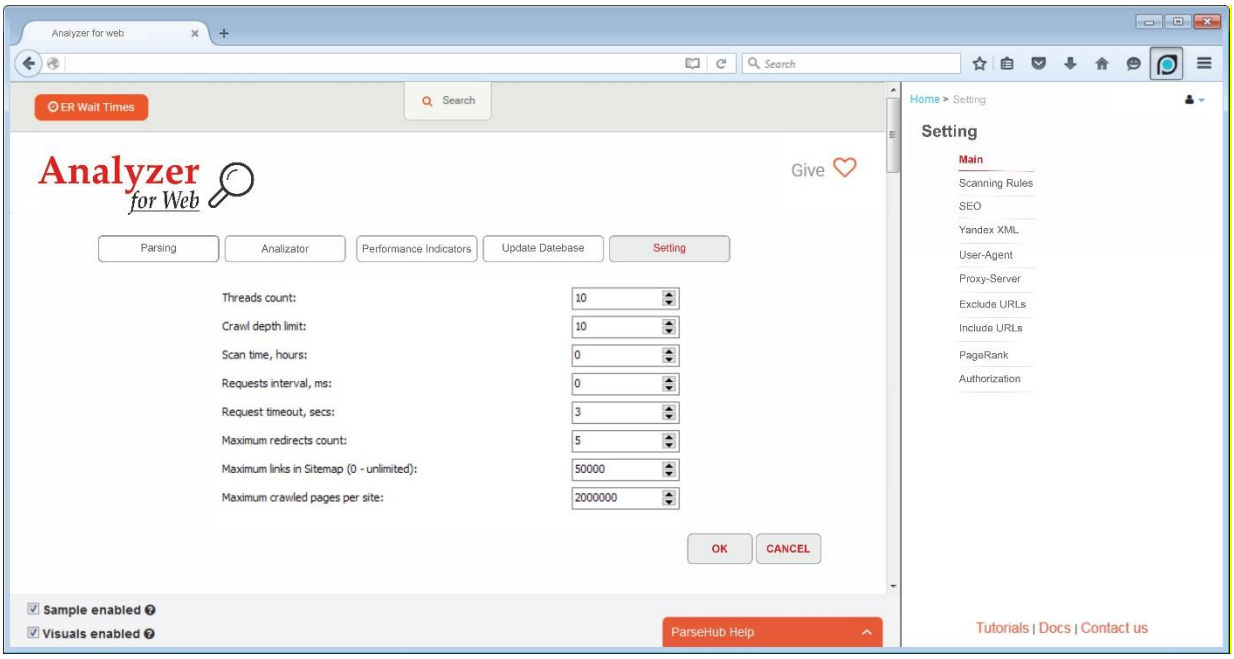


Рис. 4.15. Сторінка для налаштування параметрів для видобування інформації із веб-ресурсів

Також використовуються такі параметри налаштування: Delay between requests, Query Timeout, Maximum crawled pages.

Наступним етапом роботи із системою є визначення правил сканування. На рисунку 4.16 представлено сторінку для вибору правил сканування. Серед основних елементів управління, які подано на даній сторінці, необхідно відзначити:

1. Content types – обираються типи даних, що їх аналізатор враховує для сканування сторінок (зображення, відео, стилі, сценарії) або виключає непотрібну інформацію при аналізі.

2. Scanning Rules – налаштування зв’язані із параметрами виключення під час сканування сайту за допомогою файлу «robots.txt», посилань «nofollow» та використання директив «meta name = 'robots'» у коді сторінки веб-ресурсу.

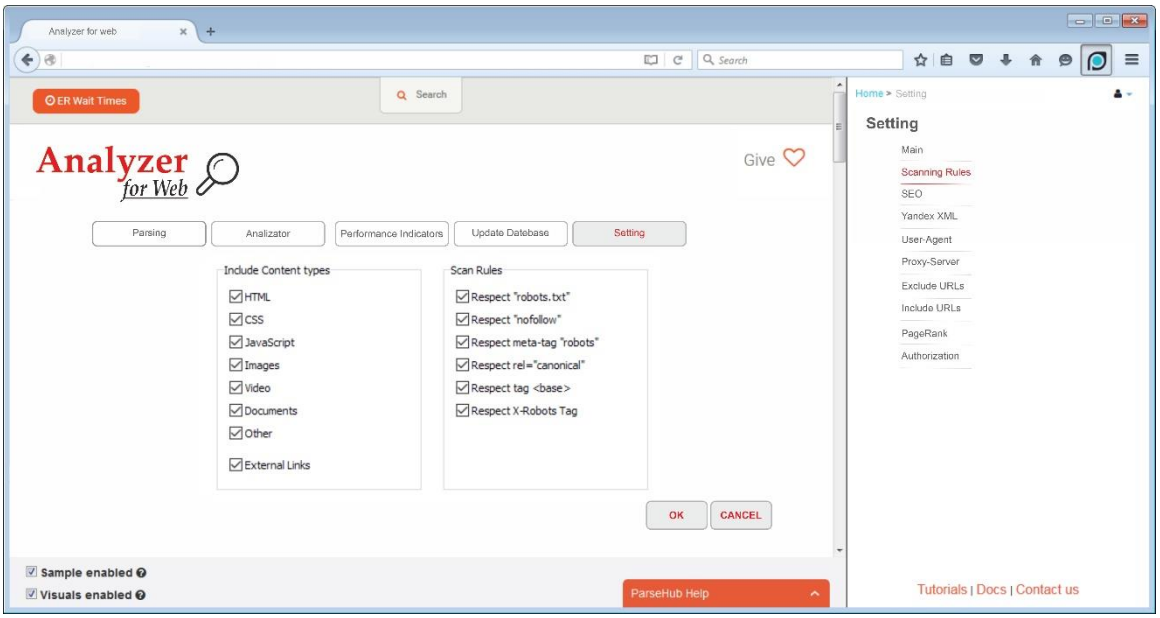


Рис. 4.16. Сторінка визначення правил сканування

Наступним етапом роботи із системою є Help. На рисунку 4.17 представлено Help для полегшення налаштування системи та подано відповіді на актуальні запитання.

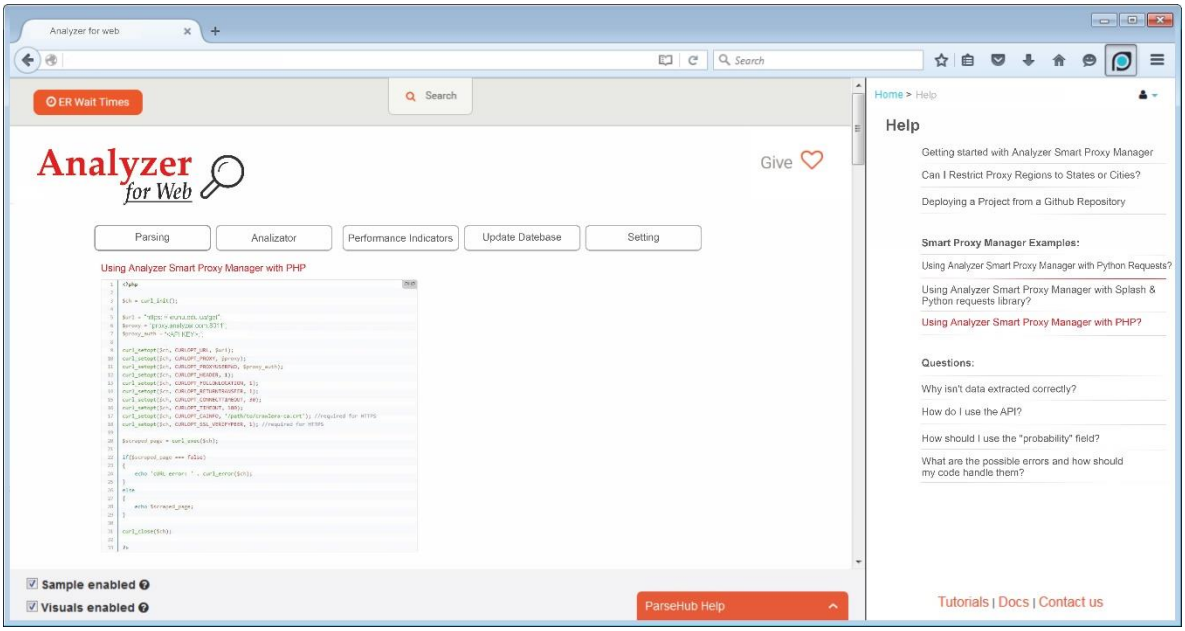


Рис. 4.17. Сторінка для оцінки ефективності роботи веб-ресурсу, виходячи із результатів структурованості контенту

Розроблена система для структурування та розпізнавання текстового контенту веб-ресурсів може використовуватися як окреме рішення або може

інтегруватися у цільові інформаційні веб-ресурси у вигляді окремого плагіна. Такий підхід дозволяє значно розширити можливості апробації та подальшого її удосконалення.

4.3. Застосування прикладної програмної системи для підтримки функціонування інформаційних веб-ресурсів.

4.3.1 Застосування прикладної програмної системи для підтримки функціонування інформаційного веб-ресурсу Західноукраїнського національного університету та дослідження його ефективності

У процесі виконання дисертаційного дослідження проведено ряд експерементальних досліджень з метою оцінки ефективності запропонованих методів та моделей. Першим інформаційним ресурсом, який досліджувався, був веб-ресурс кафедри комп'ютерних наук – http://dcs.wunu.edu.ua.

На рисунку 4.18 представлено результати роботи Web Analyzer для вказаного об'кта. На основі отриманих даних встановлено відповідність інформації, представленої на сайті, та еталоної бази даних.

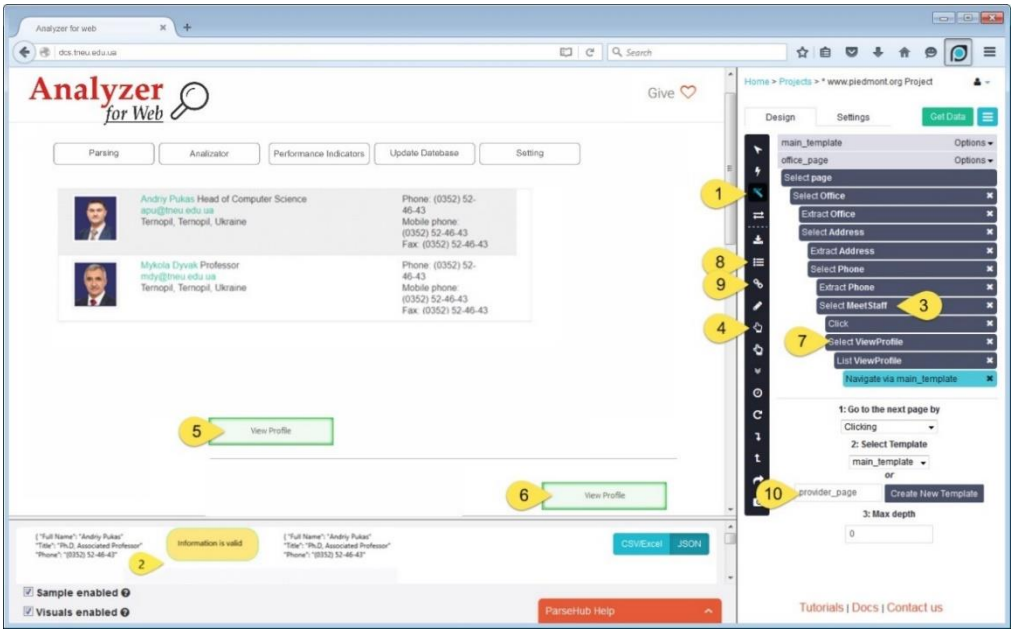


Рисунок 4.18 Результати роботи Web Analyzer із веб-ресурсом кафедри комп'ютерних наук - http://dcs.wunu.edu.ua

Наступним інформаційним ресурсом, який досліджувався, був веб-ресурс з представленою інформацією про структуру юридичного факультету. На рисунку 4.19 подано результати роботи Web Analyzer для вказаного об’кта. На основі отриманих даних встановлено ряд невідповідностей інформації, представленої на сайті, та у еталоній базі даних, а саме: невідповідність назви кафедри міжнародного права та відсутність запису у БД про юридичну клініку.

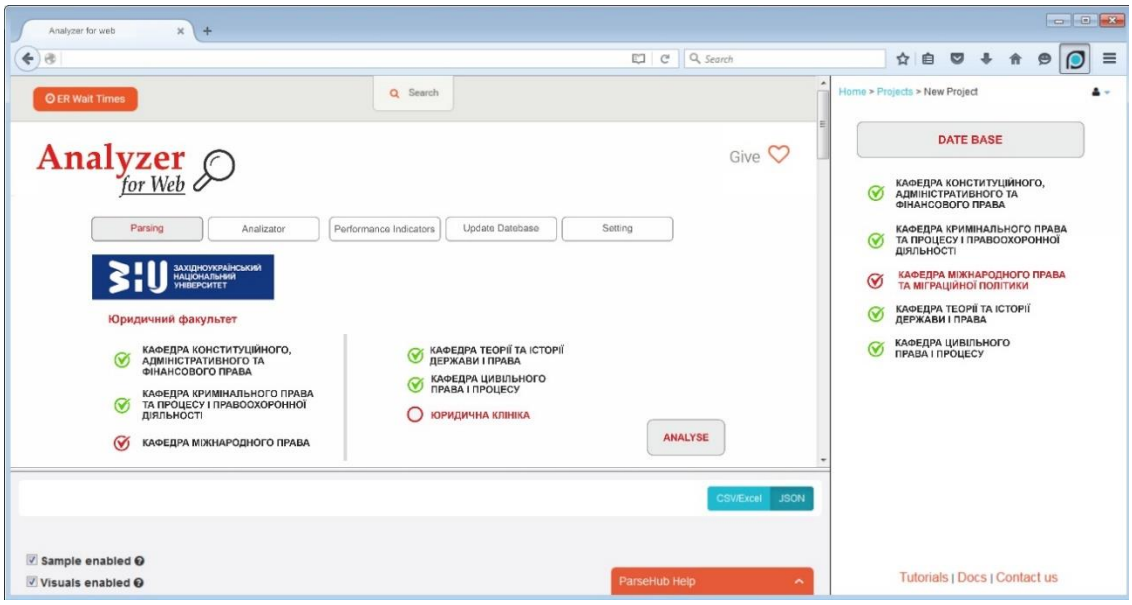


Рисунок 4.19 Результати роботи Web Analyzer із веб-ресурсом з представленою інформацією про структуру юридичного факультету

Проведені експеременти підтвердили ефективність роботи системи із веб-ресурсами інформаційно- довідкого характеру. Результати проведених досліджень, що до підвищення ефективності функціонування веб ресурсу впроваджено у організації, що підтвердженно актом (додаток 1). Застосування розробленого методу для інформаційного веб-ресурсу Західноукраїнського національного університету забезпечило зменшення часу відгуку від 5 до 20% залежно від структурованості контенту сторінок та зменшило навантаження на сервер у процедурах оновлення контенту під час SEO оптимізації.

4.3.2. Експериментальні дослідження методу моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів на прикладі Центру надання адміністративних послуг Тернопільської міської ради та ТОВ «Терногаз»

На основі запропонованого методу та моделі оцінки ефективності функціонування інформаційних веб-ресурсів вдалося визначити набір оптимальних параметрів продуктивності, враховуючи вплив структурованості контенту на ефективність функціонування системи під навантаженням. На рисунку 4.20 представлено результати аналізу показників: Time to First Byte, Database Performance, Full Page Load Time для веб-ресурсу Центру надання адміністративних послуг Тернопільської міської ради. На графіку рисунка 4.20 представлено результати моделювання для структурованого контенту з метою знаходження балансу між часом, затраченим на структурування контенту, та витратами, пов’язаними із оновленням апаратного забезпечення.

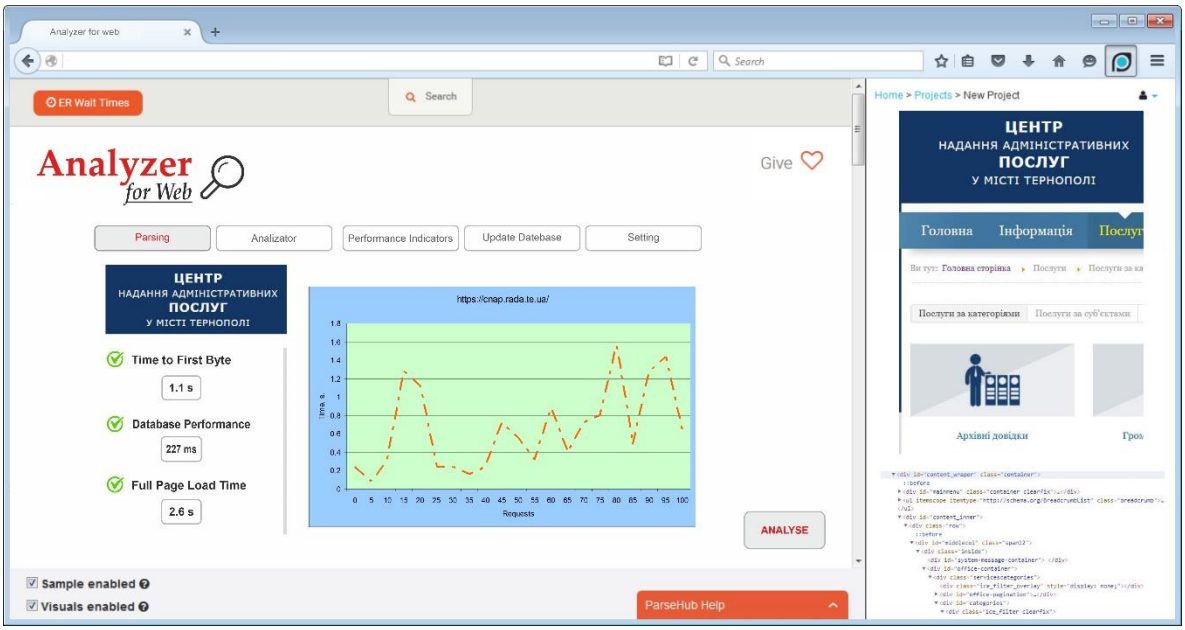


Рисунок 4.20 Результати моделювання для структурованого контенту ресурсу https://cnap.rada.te.ua/

На рисунку 4.21 подаємо есперементальні дослідження оцінки ефективності для веб-ресурсу ТОВ «Терногаз» – http://www.ternogaz.com/.

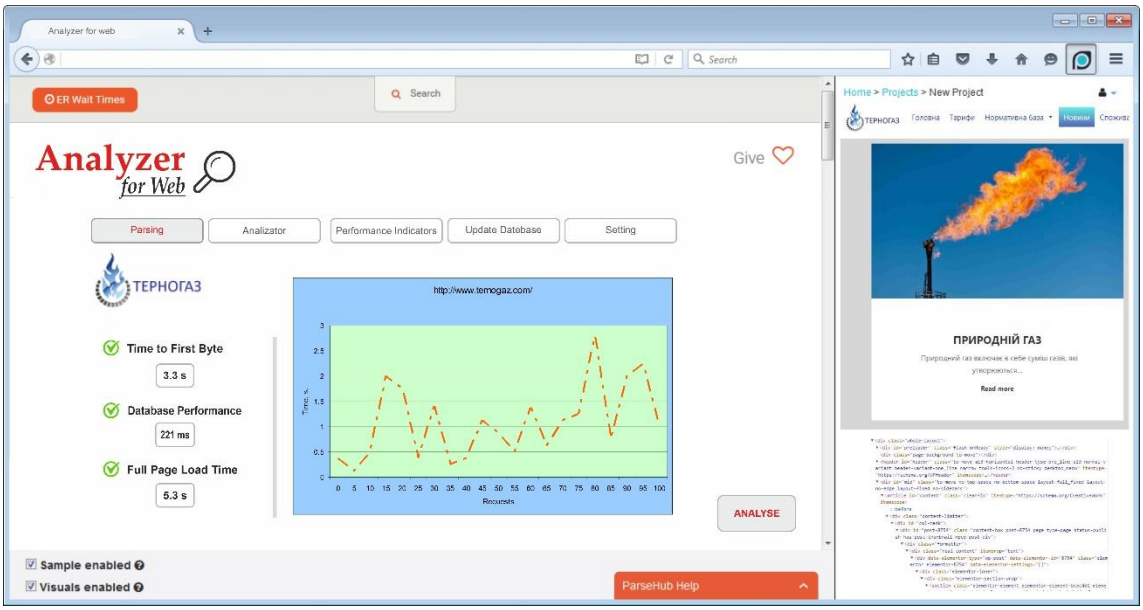


Рисунок 4.21 Результати моделювання для структурованого контенту ресурсу http://www.ternogaz.com//

Як видно із графіка, представленого на рисунку вище, інформація, що розміщується на даному сайті, є менш структурованою порівняно із веб-ресурсом ЦНАП, який аналізувався раніше. Цей висновок можна зробити виходячи із збільшення часу опрацювання запитів для користувачів веб-ресурсу. Результати проведених досліджень, що до підвищення ефективності функіонування систем впроваджено у організації, що підтвердженно актом(додаток 2-3).

Підсумовуючи проведений аналіз, можна зробити наступні висновки:

- Структурування контенту має прямий вплив на швидкість опрацювання інформаційних запитів.
- Знаходження балансу між часом, затраченим на структурування контенту, та витратами на оновлення апаратного забезпечення дозволяє сформувати правильну стратегію розвитку функіонування інформаціного ресурсу в майбутньому.

Застосування удосконаленого методу зменшило навантаження на SEO інформаційного веб-ресурсу Центр надання адміністративних послуг у місті Тернополі та ТОВ «ТЕРНОГАЗ», за оцінками експертів, на 20%.

Висновки до четвертого розділу

1. Сформульовано функціональні вимоги до прикладної системи аналізу контенту та оцінювання ефективності інформаційного веб-ресурсу. Спроектовано модулі зазначеної системи та інтерфейс користувача.

2. Розроблено архітектуру та структуру бази даних структурованого контенту для програмної системи, що виконує функції автоматизованого пошуку застарілої та некоректної інформації. Запропоновану архітектуру та узагальненні алгоритми реалізації її компонентів апробовано на ряді прикладів із веб-ресурсів навчальних закладів, чим підтверджено як достовірність результатів щодо розробки методу структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання, так і достовірність висловлених гіпотез, на яких ґрунтується запропонований підхід до побудови контент-аналізу. На тестових прикладах показано, що застосування створеного програмного забезпечення на їх основі зменшує час відгуку ресурсу від 5 до 20%.

3. Проведено апробацію розробленої прикладної програмної системи для розв’язування задачі підтримки функціонування інформаційного веб-ресурсу Західноукраїнського національного університету, зокрема, для виявлення некоректної та неактуальної інформації. Застосування розробленого методу для інформаційного веб-ресурсу Західноукраїнського національного університету забезпечило зменшення часу відгуку від 5 до 20% залежно від структурованості контенту сторінок та зменшило навантаження на сервер у процедурах оновлення контенту під час SEO оптимізації.

4. Проведено дослідження та моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів на прикладі Центру надання адміністративних послуг Тернопільської міської ради та ТОВ «Терногаз». Підтверджено ефективність та достовірність запропонованих методів. Застосування удосконаленого методу зменшило навантаження на службу SEO інформаційного веб-ресурсу Центр надання адміністративних послуг у місті Тернополі та ТОВ «ТЕРНОГАЗ», за оцінками експертів, на 20%. На тестових прикладах показано, що застосування створеного програмного забезпечення на їх основі зменшує час відгуку ресурсу від 5 до 20%.

ВИСНОВКИ

У дисертаційній роботі розв’язано актуальне науково-технічне завдання розробки математичного та програмного забезпечення для автоматизованого структурування та розпізнавання контенту із використанням засобів машинного навчання, що у сукупності забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів.

При цьому отримано такі наукові та практичні результати.

– Проаналізовано існуючі методи, засоби та технології організації інформаційних веб-ресурсів та проблеми підтримки їх функціонування. Встановлено, що ефективність функціонування інформаційних веб-ресурсів визначається рядом чинників, серед яких варто виділити продуктивність апаратного і програмного забезпечення, використовуваного джерелами, та, з іншого боку, способи організації джерела, зокрема структурованість контенту, та складність запиту. Якщо перша група характеристик вирішується у спосіб перенесення ресурсу на програмно-технічні засоби з вищою продуктивністю, то друга група вимагає від проектувальників інформаційних веб-ресурсів зусиль, пов’язаних із забезпеченням високої якості проектування та оптимізації способів організації контенту. Простіше кажучи – підвищення якості його структурування.

– Запропоновано та обґрунтовано базовий показник оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту, який визначає очікуваний час реакції ресурсу на запити. Розроблено новий метод моделювання цієї характеристики, який, на відміну від вже існуючих, дає можливість врахувати характерні для нелінійних систем особливості, такі як: залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристик. Отримані на основі запропонованого методу математичні моделі забезпечили вирішення цілого ряду практичних завдань, таких як: налаштування параметрів продуктивності інформаційного веб-ресурсу, аналіз впливу структурованості контенту, оцінка ефективності функціонування під

навантаженням, що для тестових задач забезпечило зменшення часу відгуку на 15%.

– Розроблено новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання та із використанням автоматизованого структурування контенту для виявлення недостовірної, неточної чи неактуальної інформації. Застосування розробленого методу для інформаційного веб-ресурсу Західноукраїнського національного університету забезпечило зменшення часу відгуку від 5 до 20% залежно від структурованості контенту сторінок та зменшило навантаження на сервер у процедурах оновлення контенту під час SEO оптимізації.

– Удосконалено методи автоматизованого контент-аналізу, пошуку застарілої та некоректної інформації з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, створеними як виявлення подібного контенту на інших веб-ресурсах. Застосування удосконаленого методу зменшило навантаження на SEO інформаційного веб-ресурсу Центр надання адміністративних послуг у місті Тернополі та ТОВ «Терногаз», за оцінками експертів, на 20%.

– Розроблено архітектуру прикладної програмної системи, що виконує функції автоматизованого пошуку застарілої та некоректної інформації, яка містить модулі розпізнавання і порівняння текстового контенту веб-ресурсів із елементами машинного навчання, що, у свою чергу, забезпечує підвищення ефективності функціонування прикладних програмних систем категорії інформаційні веб-ресурси. На тестових прикладах показано, що застосування створеного програмного забезпечення на їх основі зменшує час відгуку ресурсу від 5 до 20%.

– Достовірність отриманих наукових результатів підтверджено апробацією запропонованих методів та архітектур прикладних програмних систем при вирішенні прикладних задач щодо автоматизованого структурування контенту, пошуку застарілої та некоректної інформації на інформаційних веб-ресурсах та впровадженні створеної системи. Налаштування параметрів

продуктивності та аналіз впливу структурованості контенту. Створене програмне забезпечення впроваджено у Західноукраїнському національному університеті та Центр надання адміністративних послуг у місті Тернополі та ТОВ «ТЕРНОГАЗ».

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Дивак М. П., Ковбасістий А.В. Особливості побудови методу виявлення некоректної та застарілої інформації. *Сучасні комп'ютерні інформаційні технології*: Матеріали VI Всеукраїнської школи-семінару молодих вчених і студентів ACIT'2016. Тернопіль: ТНЕУ 2016. С. 119–120.

2. Melnyk A., Dyvak M., Kovbasistyi A., Brych V., Spivak I. 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), 20–24 April, 2016, Polyana-Svalyava (Zakarpattya), UKRAINE. С. 94–96. (IEEE Xplore Digital Library, DOI: 10.1109/MEMTECH.2017.7937555 P. 94–96).

3. Dyvak M., Kovbasistyi A., Stakhiv P., Lipiński P. Generalized structure of the algorithm for automated detection of non relevant and wrong information on web resources. *Journal of Applied Computer Science*. 2017. С. 23–37.

4. Dyvak M., Melnyk A., Kovbasistyi A., Turchyn L., Martsenyuk Y. System for web resources content structuring and recognizing with the machine learning elements. *Radio Electronics Computer Science Control*. Zaporizhzhia, 2018. No 3(46). P. 128–134.

5. Dyvak M., Melnyk A., Kovbasistyi A., Shcherbiak I., Huhul O. Recognition of relevance of web resource content based on analysis of semantic components. *Advanced Computer Information Technologies (ACIT'2019)*: IX International conference (Czech Republic). Ceske Budejovice, 2019. P. 297–302 (IEEE Xplore Digital Library, DOI: 10.1109/ACITT.2019.8779897).

6. Дивак М., Мельник А., Ковбасістий А., Папа О. Підхід до математичного моделювання ефективності WEB-ресурсів. *Оптико-електронні інформаційно-енергетичні технології*: *Міжнародний науково-технічний журнал*. 2019. №2 (38). С. 29–38.

7. Dyvak M., Melnyk A., Kovbasistyi A., Shevchuk R., Huhul O., Tymchyshyn V. Mathematical Modeling of the Estimation Process of Functioning Efficiency Level of Information Web-Resources. *10th International Conference on Advanced Computer Information Technologies Deggendorf*, GERMANY, 16–18 September, 2020. P. 492–496 (IEEE Xplore Digital Library, DOI: 10.1109/ACIT49673.2020.9208846).

8. Pasichnyk N. Mathematical modeling of the Website quality characteristics in dynamics / N.Pasichnyk, M.Dyvak, R.Pasichnyk. *Journal of Applied Computer Science*. Lodz: Technical University Press, 2014. Vol. 22 № 1. P. 171–183.

9. Пасічник Н., Дивак М. Метод та алгоритм побудови структур тематичних Веб-сайтів на основі онтологічного підходу. *Наукові праці Донецького національного технічного університету*: серія Інформатика, кібернетика та обчислювальна техніка, 2012. Вип 15 (203). С. 184–189.

10. Пасічник Н. Метод формування онтологічного контенту на основі аналізу інформації спеціалізованих Веб-сайтів. *Вісник Хмельницького національного університету:* серія Технічні науки. 2012. Вип. 5. С. 241–244.

11. Pasichnyk N., Dyvak M. Formalism in the quality site creating problem. *Naukovi pratsi DonNTY*: ser. Informatyka, kibernetyka ta obchysliuvalna tekhnika. 2011, Vol. 14(188). P. 325–329 (in Ukrainian).

12. Pasichnyk N., Dyvak M. Matrix the method and algorithm of construction of the content websites structures based on the ontological approach. *Naukovi pratsi DonNTY*: ser. Informatyka, kibernetyka ta obchysliuvalna tekhnika. 2012. Vol. 15. P. 184–189 (in Ukrainian).

13. Pasichnyk N., Method of forming an ontological content, based on analysis of information at specialized Web-sites. *Visnyk HNU*: ser. Tekhnichni nauky. 2012. Vol. 5. P. 241–244 (in Ukrainian).

14. Pasichnyk N., Dyvak M. Mathematical model of traffic dynamics of the specialized websites and methods of its identification. *Induktyvne modeliuvannia skladnykh system*: Zb. nauk. pr. 2013. Vol. 5. P. 237–247 (in Ukrainian).

15. Dyvak M., Pasichnyk R., Pasichnyk N. Identification and modeling of limiting factors systems. *Proceedings of the 2016 IEEE First International Conference on Data Stream Mining & Processing* (DSMP), 23–27 Aug. 2016. Lviv, 2016. P. 336–340.

16. The structure of the site. Creation and development of categorization. URL: http://seo-for-ucoz.com/load/podgotovka_k_prodvizheniyu/struktura_sajta/1-1-0-4

17. Analysis of site structure. URL: http://www.web-patrol.net/audit-site-struktur.html

18. Duncan Temple Lang. XML: Tools for parsing and generating XML within R and S-Plus. URL: http://CRAN.R-project.org/package=XML

19. Duncan Temple Lang. RCurl: General network (HTTP/FTP/...) client interface for R. URL: http://CRAN.R-project.org/package=RCurl

20. Xu Z., Wei X., Liu Y., Mei L., Hu C., Choo K. K. R., Sugumaran V. Building the search pattern of web users using conceptual semantic space model. *International Journal of Web and Grid Services*. 2016. Vol. 12 (3). P. 328–347.

21. Dyvak M., Pasichnyk N. Formalism in setting the task of creating a quality website. *Scientific papers of Donetsk National Technical University*: ser. Informatics, Cybernetics and Computer tehnika. 2011. Vol. 14 (188). P. 325–329.

22. Dyvak M., Pasichnyk N. The method and algorithm for constructing structures themed Web sites based on ontological approach. *Scientific papers of Donetsk National Technical University*: ser. Informatics, Cybernetics and Computer tehnika. 2012. Vol. 14 (188). P. 310–315.

23. Auer S., Domingue J. Block Chain Technologies & The Semantic Web: A Framework for Symbiotic Development. Technical report. Bonn: University of Bonn, 2016.

24. Weare C., Lin W.Y. Content Analysis of the World Wide Web: Opportunities and Challenges. *Social Science Computer Review*. 2002. Vol. 18 (272).

25. Xu Z., Wei X., Liu Y., Mei L., Hu C., Choo K. K. R., Sugumaran V. Building the search pattern of web users using conceptual semantic space model. *International Journal of Web and Grid Services*. 2016. Vol. 12 (3). P. 328–347.

26. Dyvak M., Pasichnyk N. Formalism in setting the task of creating a quality website. *Scientific papers of Donetsk National Technical University*: ser. Informatics, Cybernetics and Computer tehnika. 2011. Vol. 14 (188). P. 325–329.

27. Dyvak M., Pasichnyk N. The method and algorithm for constructing structures themed Web sites based on ontological approach. *Scientific papers of Donetsk National Technical University*: ser. Informatics, Cybernetics and Computer tehnika. 2012. Vol. 14 (188). P. 310–315.

28. Auer S., Domingue J. Block Chain Technologies & The Semantic Web: A Framework for Symbiotic Development. Technical report. Bonn: University of Bonn, 2016.

29. Weare C., Lin W.Y. Content Analysis of the World Wide Web: Opportunities and Challenges. *Social Science Computer Review*. 2002. Vol. 18 (272).

30. Abernethy J., Bartlett P. L., Rakhlin A. & Tewari A. Optimal strategies and minimax lower bounds for online convex games. *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008.

31. Barber D. Bayesian reasoning and machine learning. Cambridge: Cambridge University Press, 2012. Bartlett P., Bousquet O. & Mendelson S. Local rademacher complexities. *Annals of Statistics*. 2005. Vol. 33(4). P. 1497–1537.

32. Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning.* 2009. Vol. 2(1). P. 1–127.

33. Le Q. V., Ranzato M.-A., Monga R., Devin M., Corrado G., Chen K., Dean J. & Ng A. Y. Building high-level features using large scale unsupervised learning. *International Conference on Machine Learning* (ICML), 2012.

34. Mayring P. Qualitative content analysis. *Forum: Qualitative Social Research*. 2000. Vol. 1(2). Retrieved July 28, 2008. URL: http://217.160.35.246/fqs-texte/2-00/2- 00mayring-e.pdf.

35. Gerbic P. & Stacey E. (2005) A purposive approach to content analysis: designing analytical frameworks. *Internet and Higher Education*. 2005. Vol. 8. P. 45–59.

36. Tools for parsing in the work of an SEO specialist. URL: https://netpeak.net/ru/blog/instrumenty-dlya-parsinga-v-rabote-seo-spetsialista/

37. Abernethy J., Bartlett P. L., Rakhlin A. & Tewari A. Optimal strategies and minimax lower bounds for online convex games. *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008.

38. Barber D. Bayesian reasoning and machine learning. Cambridge: Cambridge University Press, 2012. Bartlett P., Bousquet O. & Mendelson S. Local rademacher complexities. *Annals of Statistics*. 2005. Vol. 33(4). P. 1497–1537.

39. Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*. 2009. Vol. 2(1). P. 1–127.

40. Le Q. V., Ranzato M.-A., Monga R., Devin M., Corrado G., Chen K., Dean J. & Ng A. Y. Building high-level features using large scale unsupervised learning. *International Conference on Machine Learning* (ICML), 2012.

41. Mayring P. Qualitative content analysis. *Forum: Qualitative Social Research*. 2000. Vol. 1(2). Retrieved July 28, 2008. URL: http://217.160.35.246/fqs-texte/2-00/2- 00mayring-e.pdf.

42. Gerbic P. & Stacey E. (2005) A purposive approach to content analysis: designing analytical frameworks. *Internet and Higher Education*. 2005. Vol. 8. P. 45–59.

43. vTools for parsing in the work of an SEO specialist. URL: https://netpeak.net/ru/blog/instrumenty-dlya-parsinga-v-rabote-seo-spetsialista/.

44. Weare C., Lin W.Y., Content Analysis of the World Wide Web: Opportunities and Challenges. *Social Science Computer Review*. 2002. Vol. 18(272).

45. Singh N., Baack D.W. Web Site Adaptation: A Cross-cultural Comparison of U.S. and Mexican Web sites. *Journal of Computer-mediated Communication*. 2004. Vol. 9(4).

46. Aone C. & Bennett S.W. Applyingf machine learning to anaphora resolution. In Wermter S., Riloff E. & Scheler G. (Eds.). Connectionist, statistical and symbolic approaches to learning for natural language processing. Berlin: Springer-Verlag, 1996. P. 302–314

47. Ghahramani Z. Unsupervised learning algorithms are designed to extract structure from data, 2008. P. 1–8. IOS Press.

48. Harrington P. Machine Learning in Action. New York: Manning Publications Co., Shelter Island, 2012 (ISBN 9781617290183).

49. Xu L., Holger H. Hoos, Leyton-Brown K. Hydra: Automatically configuring algorithms selection. *Twenty-Fourth Conference of the Association for the Advancement of Artificial Intelligence* (AAAI-10), 2010. P. 210–216.

50. Witten I., Frank E., Hall M. Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition. San Mateo: Morgan Kaufmann, CA, 2011.

51. Smola A., S.V.N. Vishwanathan: Introduction to Machine Learning, Cambridge: Cambridge University Press, 2008; eBook (October 1, 2010). P. 234.

52. Ayodele T. O. Types of machine learning algorithms. *New Advances in Machine Learning*, Rijeka: InTech, 2010.

53. Tong H., Lim K.S. Threshold autoregression, limit cycles, and cyclical data (with discussion). *Journal of Royal Statistical Society*, Series B. 1980. Vol.42. P. 245–292. Maxwell J. Clerk. A Treatise on Electricity and Magnetism, 3rd ed. Oxford: Clarendon, 1892, Vol. 2. P. 68–73.

54. Tong H. Non-linear time series: a dynamical approach. Oxford: Clarendon Press Oxford, 1990. 564 p.

55. Pasichnyk N., Melnyk A., Dobrovolska N. Management the Website Attendance Based on the Projected Traffic. *The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM):* Proceedings of the international conference, 19-23 Feb. 2013. Polyana-Svalyava (Zakarpattya), 2013. P. 277.

56. Ivanov A. Mathematical models of basic processes of information web portal functioning. *Systems and Means of Informatics*. 2010. Volume 20, Issue 1. P. 106–132.

57. Pasichnyk R., Melnyk A. Modeling of Cognitive Processes for Bio-Technical Systems. *Modern Problems of Radio Engineering, Telecommunications and Computer Science*: Proceedings of the International Conference. TCSET'2008. Lviv-Slavsko, 2008. P. 27–28.

58. Manhas Dr. A Study of Factors Affecting Websites Page Loading Speed for Efficient Web Performance. *International Journal of Computer Sciences and Engineering*. 2013. Vol. 1 (3). P. 32–35.

59. Bartuskova A., Soukal I. The Novel Approach to Organization and Navigation by Using all Organization Schemes Simultaneously. *Lecture Notes in Business Information Processing*. 2016. Springer. 261. P. 99–106.

60. Pasichnyk R., Melnyk A. Modeling of effective studies in Adaptive Educational Systems. *The Expirence of Designing and Application of CAD Systems in Microelectronics*: Proceedings of the International Conference. CADSM '2009. Lviv-Polyana, 2009. P. 248–250.

61. Pasichnyk R., Melnyk A. Modeling of effective studies in Adaptive Educational Systems. *The Expirence of Designing and Application of CAD Systems in Microelectronics*: Proceedings of the International Conference. CADSM '2009. Lviv-Polyana, 2009. P. 248–250.

62. Data Mining Techniques. URL: https://www.javatpoint.com/data-mining-techniques

63. Берко А. Ю. Системи електронної контент-комерції: монографія / А. Ю. Берко, В. А. Висоцька, В. В. Пасічник. Львів: Вид-во Нац. ун-ту «Львівська політехніка», 2009. 612 с.

64. Хорошилова Т. Зміст методики «контент-аналіз». *Прикладна лінгвістика*. URL: http://studentstpl.ucoz.ru/publ/teorija_vozdejstvija/metodika_kontent_ analiza/ soderzhanie_ metodiki_kontent_analiz/12-1-0-116.

65. Математична лінґвістика: навч. посібник. Книга 1. Квантитативна лінгвістика /В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. Львів: «Новий світ -2000», 2012. 359 с.

66. Григорьев С. Проведение контентанализа: навч. посібник. URL: http://www.psyfactor.org/lib/k-a2.htm

67. Манаєв О. Контент-аналіз як метод дослідження. *Псі-фактор*. URL: http://psyfactor.org/lib/content-analysis3.htm

68. Аналіз статистики: Методи багатовимірної статистики. URL: http://christsocio.info/content/view/492/102/

69. The Content Management Possibilities Poster. URL: http://metatorial.com/pagea.asp? id=poster.

70. Methods based on ontologies for information resources processing: Monograph / V. Lytvyn, V. Vysotska, L. Chyrun, D. Dosyn. *LAP Lambert Academic Publishing*. Saarbrücken, 2016. 324 c. (ISBN-13: 978-3-659-89905-8, ISBN10: 3659899054, EAN: 9783659899058).

71. Berko A. Features of information resources processing in electronic content commerce / A. Berko, V. Vysotska, L. Chyrun. *Applied Computer Science. ACS journal*. 2014. Vol. 10, Number 2. P. 5–19. (ISSN 2353-6977 (Online), ISSN 1895-3735 (Print).

72. Vysotska V., Chyrun L. Web Content Processing Method for Electronic Business Systems. *International Journal of Computers & Technology*. 2013. Vol. 12, No 2. December. P. 3211–3220. (ISSN 2277-3061) URL: http://cirworld.org/journals/index.php/ijct/article/view/3299. Impacr factor 1,532. (Index Copernicus, NASA ADS, DOAJ, Google Scholar, Eyesource, EBSCO, CiteSeer, UlrichWeb, ScientificCommons, ProQuest CSA Technology Research Database).

73. Чирун Л., Висоцька В. Застосування контент-аналізу текстової інформації в системах електронної комерції. *Вісник Національного університету «Львівська політехніка»:* серія Інформаційні системи та мережі. 2010. № 689. С. 332–347.

74. Almeida F., Santos J. D., & Monteiro J. A. (2013). ECommerce business models with the context of Web 3.0 paradigm. *International Journal of Advanced Information Technology*. 2013. Vol. 3(6). P. 1–12.

75. Aghaei S., Nematbakhsh M.A. and Farsani H.K. Evolution of the World Wide Web: From Web 1.0 to Web 4.0. *Int'l J. Web & Semantic Technology*. 2012. vol. 3, no. 1. P. 1–10.

76. Web 1.0 vs Web 2.0 vs Web 3.0 vs Web 4.0 vs Web 5.0 – A bird's eye on the evolution and definition. URL: https://flatworldbusiness.wordpress.com/flat-education/previously/web-1-0-vs-web-2-0-vs-web-3-0-a-bird-eye-onthe-definition.

77. Divide the Web Timeline in nine epochs. URL: http://www.web3.lu/divide-the-web-timeline-in-nine-epochs.

78. Calacanis J. Web 3.0, the «official» definition. URL: http://calacanis.com/2007/10/03/web-3-0-the-official-definition.

79. URL: http://www.bourabai.kz/dbt/web/evolution.htm (дата звернення: 14.05.2018).

80. Тренды для веб разработки в 2017. URL: https://artjoker.ua/ru/blog/trendy-dlya-veb-razrabotki-v-2017.

81. Тренды веб-разработки 2018. URL: http://merehead.com/blog-ru/web-development-trends-2018.

82. URL: https://pedagogy.bdpu.org/wp-content/uploads/2020/05/5.pdf

83. Su C.-J., Huang S.-F. Real-time big data analytics for hard disk drive predictive maintenance. Comput. Electr. Eng. 2018. P. 71, 93–101. [CrossRef].

84. Micheloni R. Solid-State Drive (SSD): A Nonvolatile Storage System; IEEE Xplore. Piscataway, NJ, 2017. Vol. 105. P. 583–588.

85. Spinelli A.S., Compagnoni C.M., Lacaita A.L. Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices. *Computers*. 2017. Vol. 6. P. 16. [CrossRef]. Ielmini D. Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks. *Microelectron. Eng*. 2018. Vol. 190. P. 44–53. [CrossRef]

86. Asana I.M.D.P., Wiguna I.K.A.G., Atmaja K.J., Sanjaya I.P.A. FP-Growth Implementation in Frequent Itemset Mining for Consumer Shopping Pattern Analysis Application. Mantik J. 2020, 4, P. 2063–2070.

87. Duriez B., Chilaka S., Bercher J.-F., Hercul E. and Prioleau M.-N. Replication dynamics of individual loci in single living cells reveal changes in the degree of replication stochasticity through S phase. *Nucleic Acids Research*. 2019. vol. 47. P. 5155-5169.

88. Kraus M., Feuerriegel S., Oztekin A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*. 2020. Vol. 281(3). P. 628–641. URL: https://dx.doi.org/10.1016/j.ejor.2019.09.018.

89. Matsyi O. Approaches to solving basic problems of closed routes. *IEEE Proceedings of 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering* (TCSET-2020). Lviv-Slavske, 2020. February 25–29. P. 69–72. (Scopus).

90. Saurkar A., Pathare K., Gode S. An Overview On Web Scraping Techniques And Tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*. 2018. Vol. 4(4). P. 363–367.

91. Thomas DM, Mathur S. Data analysis by web scraping using python. *3rd International conference on Electronics, Communication and Aerospace Technology* (ICECA). 2019. P. 450–454. URL: https://ieeexplore.ieee.org/abstract/document/8822022.

92. Gutiérrez S., Torres N., Domínguez N. Análisis con «big data» de las respuestas de los hoteles en TripAdvisor. Esic Mark. 2018, P. 160, 339–378.

93. Vania C., Kementchedjhieva Y., Søgaard A., Lopez A. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). Hong Kong, November, 2019. P. 1105–1116.

94. Wang D., Eisner J. 2018. Synthetic data made to order: The case of parsing. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Brussels, October-November, 2018. P. 1325–1337,.

95. d'Amato C. Machine finding out for the Semantic Web: Classes learnt and subsequent research instructions. *Semantic Web*. 2020. Vol. 11, 1. P. 195–203.

96. Guha R., Brickley D., Macbeth S. Schema.org: evolution of structured files on the on-line. *Commun. ACM*. 2016. Vol. 59, 2. P. 44–51. URL:https://doi.org/10.1145/2844544

97. Haller A., Janowicz K., Cox S., Phuoc D., Taylor K., Lefrancois M. (Eds.). Semantic Sensor Network Ontology. *W3C Recommendation 19 October*, 2017. URL:http://www.w3.org/TR/vocabssn/.

98. Hammar K. et al. Easy research questions relating ontology uncover patterns. Ontology Engineering with Ontology Salvage Patterns—Foundations and Functions. Hitzler P., Gangemi A., Janowicz K., Krisnadhi A., Presutti V. (Eds.). Studies on the Semantic Web 25. IOS Press, 2016. P. 189–198.

99. Hitzler P., Bianchi F., Ebrahimi M., Sarker M. Neural-symbolic integration and the Semantic Web. *Semantic Web.* 2020. Vol. 11, 1. P. 3–11.

100. Jacksi K., Abass S.M. Development History Of The World Wide Web. *International Journal of Scientific & Technology Research* (IJSTR). 2019. Vol. 8, No 09. P.75–79.

101. Naik U., Shivalingaiah D. Comparative study of Web1.0, Web 2.0, Web 3.0. *6 th International CALIBER-2008* (Allahabad, 29 February-1 March 1 2008). Allahabad, 2008. P. 499–507. URL: http://hdl.handle.net/1944/1285.

102. Lal M. Web 3.0 in education and research. *BVICAM's International Journal of Information Technology*. 2011. Vol. 3, No. 2. P. 335–340.

103. Chisega-Negrila A.M. Education in Web 3.0. *Journal of Advanced Distributed Learning Technology*. 2013. Vol.3, No 1. P. 50–58.

104. Potapchuk O. Application of web-technologies in the educational process of higher educational institutions of Ukraine. *Journal of Education, Health and Sport.* 2018. Vol 8, No 2. P. 235–242. eISSN 2391-8306. DOI http://dx.doi.org/10.5281/zenodo.1175249.

105. Habibi A., Mukminin A., Pratama R., Harja H. Predicting Factors Affecting Intention to use Web 2.0 in Learning: Evidence from Science Education. *Journal of Baltic Science Education*. 2019. Vol. 18, No. 4. P. 595–606.

106. Bernstein A., Hendler J., Noy N. A brand fresh question on the Semantic Web. *Commun. ACM.* 2016. Vol. 59, 9. P. 35–37.

107. d'Amato C. Machine finding out for the Semantic Web: Classes learnt and subsequent research instructions. *Semantic Web*. 2020. Vol. 11, 1. P. 195–203.

108. Guha R., Brickley D., Macbeth S. 2016. Schema.org: evolution of structured files on the on-line. *Commun.* 2016. ACM Vol. 59, 2. P. 44–51. URL: https://doi.org/10.1145/2844544

109. Hitzler P., Bianchi F., Ebrahimi M., Sarker M. Neural-symbolic integration and the Semantic Web. *Semantic Web*. 2020. Vol. 11, 1. P. 3–11.

110. Hyvönen E. The utilization of the Semantic Web in digital humanities: Shift from files publishing to files-prognosis and serendipidous knowledge discovery. *Semantic Web*. 2020. Vol. 11, 1. P. 187–193.

111. Lecue F. On the feature of knowlege graphs in explainable AI. *Semantic Web*. 2020. Vol. 11, 1. P. 41–51.

112. Martinez-Rodriguez J., Hogan A., Lopez-Arevalo I. Info extraction meets the Semantic Web: A Peek. *Semantic Web*. 2020. Vol. 11, 2. P. 255–335.

113. Thieblin E., Haemmerle O., Hernandez N., Santos C. Peek on advanced ontology matching. *Semantic Web*. 2020. Vol. 11, 2. P. 689–727.

114. Selamat S. A. M., Prakoonwit S., Khan W. A review of data mining in knowledge management: applications/findings for transportation of small and medium enterprises». *SN Applied Sciences.* 2020. вип. 2, вип. 5, С. 1–15.

115. Velásquez-Pérez T., Camargo-Pérez J. A., Quintero-Quintero E. L. Application of data mining as a tool in computer science». *Journal of Physics: Conference Series*. 2020. Вип. 1587, вип. 1. С. 012018.

116. Sarin P., Kar A. K., Ilavarasan V. P. Exploring engagement among mobile app developers–Insights from mining big data in user generated content. *Journal of Advances in Management Research*, 2021.

117. Kollmann T. Grundlagen des Web 1.0, Web 2.0, Web 3.0 und Web 4.0. *Handbuch Digitale Wirtschaft*, Springer, 2020. С. 133–155.

118. Negrila M. C. Impact of web 3.0 on the evolution of learning. *Conference proceedings of eLearning and Software for Education* (eLSE). 2016. Вип. 01. С. 58–62.

119. Rhayem M. B., Mhiri A., Gargouri F. Semantic web technologies for the internet of things: Systematic literature review. *Internet of Things*. 2020. С. 100206.

120. Kasemsap K. Software as a service, Semantic Web, and big data: Theories and applications. *Research Anthology on Recent Trends, Tools, and Implications of Computer Programming*, IGI Global, 2021. С. 1179–1201.

121. Firat E. A., Firat S. Web 3.0 in learning environments: a systematic review. *Turkish Online Journal of Distance Education*. 2020. Вип. 22, вип. 1. С. 148–169.

122. Pereira P., Araujo J., Torquato M., Dantas J., Melo C., Maciel P. Stochastic performance model for web server capacity planning in fog computing. *The Journal of Supercomputing*. 2020. Вип. 76, вип. 12. С. 9533–9557.

123. Гайворонская С., Шурчкова Ю. Методические подходы к оценке эффективности маркетинговых коммуникаций в сети Интернет. *Современная экономика: проблемы и решения.* 2015. № 12(72). С. 109–123. DOI: 10.17308/meps.2015.12/1355

124. Calameo 2020. Top 5 digital publishing strategy trends for 2021. Cited 19.11.2020. URL: https://blog.calameo.com/6375/top-5-digital-publishing-strategy-trends-for-2021/

125. Cozmiuc C. How to Measure the Success of Your SEO Efforts. Cited 21.09.2020. URL: https://cognitiveseo.com/blog/15516/measure-seoefforts/#:~:text=Organic%20click%2Dthrough%20rate%20is,the%20search%20result%20has%20 received

126. Dean B. Link Building for SEO: The Definitive Guide. Cited 6.11.2020. URL: https://backlinko.com/link-building

127. Papagiannis N. Effective SEO and Content Marketing. Wiley Online Library, 2020.

128. Sharma M., Joshi S. Online advertisement using web analytics software: a comparison using AHP method. *International Journal of Business Analytics* (IJBAN). 2020. Вип. 7, вип. 2. С. 13–33.

129. Kammar J., Rafi M. Survey on Web Analytics. *Journal of Web Engineering & Technology*. 2021. Вип. 7, вип. 3. С. 20–24.

130. Akhigbe I. UcEF for Semantic IR: An Integrated Context-Based Web Analytics Method. *Advanced Concepts, Methods, and Applications in Semantic Computing*, IGI Global, 2021. P. 190–217.

131. Kumar V., Ogunmola G. A. Web analytics for knowledge creation: a systematic review of tools, techniques, and practices. *International Journal of Cyber Behavior, Psychology and Learning* (IJCBPL). 2020. Вип. 10, вип. 1. С. 1–14.

132. Neal D. Web Analytics: Analyzing Users of a University Website. *PhD Thesis*, Northern Kentucky University, 2020.

133. Almeida F. L. Concept and dimensions of web 4.0. *International Journal of Computers & Technology*. 2017. Vol. 16(7). P. 7040–7046.

134. Demartini C. & Benussi L. Do Web 4.0 and industry 4.0 imply education X. 0?. *IT Professional*. 2017. Vol. 19(3). P. 4–7.

135. George G. & Lal A. M. Review of ontology-based recommender systems in e-learning. *Computers & Education*. 2019. Vol. 142. P. 103642

136. Duan Y., Edwards J.S., Dwivedi Y.K. Artificial intelligence for decision making in the era of BigData – Evolution, challenges and research agenda. Int. *J. Inf. Manag*. 2019. Vol. 48. P. 63–71.

137. Subasi A. Practical Machine Learning for Data Analysis Using Python. Academic Press, 2020.

138. Raschka S., Patterson J., Nolet C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*. 2020, Vol. 11. P.193.

139. Pedregosa F. et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011. Vol. 12. P. 2825−2830.

140. Djosic N., Nokovic B., Sharieh S. Machine Learning in Action: Securing IAM API by Risk Authentication Decision Engine. *CNS '20. IEEE* (Jun 2020). URL: https://doi.org/10.1109/CNS48642.2020.9162317

ДОДАТКИ

ЗАТВЕРДЖУЮ
Начальник відділу «Центр надання адміністративних послуг»
Тернопільської міської ради
_____ ПАНИЧЕВА І. Є.
«04» листопада 2020 р.

АКТ

№ 1024 від 04.11.2020

про впровадження результатів дисертаційної роботи
Ковбасістого Андрія Вікторовича
«Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості»

У процесі розробки та впровадження прикладного програмного забезпечення центру надання адміністративних послуг Тернопільської міської ради використано результати дисертаційної роботи здобувача кафедри комп'ютерних наук Західноукраїнського національного університету Ковбасістого Андрія Вікторовича, а саме:

— метод моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту, який дає можливість врахувати характерні для нелінійних систем особливості, такі як залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристик, налаштувати параметри продуктивності, оцінити ефективність функціонування ресурсу під навантаженням;

— новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання, який за рахунок уведення формального опису контенту та компонентів машинного навчання, забезпечує автоматизоване структурування контенту, виявлення в структурі контенту недостовірної, неточної чи неактуальної інформації;

— архітектуру прикладних програмних систем, що виконують функції автоматизованого пошуку застарілої та некоректної інформації, які містять модулі розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання.

Розроблені автором методи безпосередньо впроваджені в ЦНАП, що у сукупності забезпечило підвищення ефективності функціонування прикладної програмної системи за рахунок скорочення часу виконання запитів інформаційного веб-ресурсу.

Заступник начальника відділу
«Центр надання адміністративних послуг» А.В. ДЕМАКОВА

Завідувач сектору надання послуг
виконавчих органів ради відділу
«Центр надання адміністративних послуг»,
Адміністратор О.О. ДУДЧЕНКО

Комерційний директор
ТОВ "ТЕРНОГАЗ"
Галина ГОЛОЮХА
N 241 від 30.10.

АКТ

про впровадження результатів дисертаційної роботи
Ковбасістого Андрія Вікторовича
«Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості»

У процесі розробки та впровадження прикладного програмного забезпечення для ТОВТЕРНОГАЗ використано результати дисертаційної роботи здобувача ступеня доктора філософії Західноукраїнського національного університету Ковбасістого Андрія Вікторовича, а саме:

- метод моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів залежно від структурованості контенту, який дає можливість врахувати характерні для нелінійних систем особливості, такі як циклічність процесів, стрибкоподібні зміни характеристик,а також налаштувати параметри продуктивності, оцінити ефективність функціонування ресурсу під навантаженням;
- новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання, який за рахунок уведення формального опису контенту та компонентів машинного навчання, забезпечує автоматизоване структурування контенту, виявлення в структурі контенту недостовірної, неточної чи неактуальної інформації;

Розроблені автором методи безпосередньо впроваджені в ТОВ ТЕРНОГАЗ, що у сукупності забезпечило підвищення ефективностіфункціонування онлайн система “Терногаз” за рахунок скорочення часу виконання запитів для користувачів.

Інженер-програміст Олександр КАЛИТА

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Західноукраїнського
національного університету
д.е.н., проф. Зеновій ЗАДОРОЖНИЙ
_____
«__»_____2020 р.

АКТ

про впровадження результатів дисертаційної роботи
Ковбасістого Андрія Вікторовича
«Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості»

У процесі розробки та впровадження прикладного програмного забезпечення для веб-сайту Західноукраїнського національного університету використано результати дисертаційної роботи здобувача ступеня доктора філософії Ковбасістого Андрія Вікторовича, а саме:

- метод моделювання базового показника оцінки ефективності функціонування інформаційних веб-ресурсів, залежно від структурованості контенту, який дає можливість врахувати характерні для нелінійних систем особливості, циклічність процесів, стрибкоподібні зміни характеристик, а також налаштувати параметри продуктивності, оцінити ефективність функціонування ресурсу під навантаженням;
- новий метод структурування, розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання, який за рахунок уведення формального опису контенту та компонентів машинного навчання, забезпечує автоматизоване структурування контенту, виявлення в структурі контенту недостовірної, неточної чи неактуальної інформації;
- удосконалені методи автоматизованого контент-аналізу, пошуку застарілої та некоректної інформації з процедурами перетворення інформації із інформаційних веб-ресурсів у формат, придатний для порівняння з базами даних, створеними у спосіб виявлення подібного контенту на інших веб-ресурсах.

Розроблені автором методи безпосередньо впроваджені в Західноукраїнському національному університеті, що у сукупності забезпечило підвищення ефективності функціонування веб сайту університету за рахунок скорочення часу виконання запитів інформаційного веб-ресурсу та виявлення недостовірної, неточної чи неактуальної інформації.

Директор навчально-наукового
центру інформаційних технологій Ігор РОМАНЕЦЬ

ЗУНУ
№ 126-20/1778 від 11.11.2020

ЗАТВЕРДЖУЮ
Перший проректор
Західноукраїнського
національного університету
к. ф.-м. н. Микола ШИНКАРИК
_____
« » _____2020 р.

АКТ

про впровадження в навчальний процес Західноукраїнського національного університету результатів дисертаційної роботи
Ковбасістого Андрія Вікторовича
«Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості»

Даний акт складений про те, що результати дисертаційної роботи здобувача ступеня доктора філософії Ковбасістого Андрія Вікторовича на тему: «Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості» використані в освітньому процесі факультету комп'ютерних інформаційних технологій Західноукраїнського національного університету для студентів спеціальності «Інженерія програмного забезпечення».

При викладанні дисциплін «Веб-програмування» та «Веб-онтології» розглядаються методи побудови архітектури прикладних програмних систем, що виконують функції автоматизованого пошуку застарілої та некоректної інформації, які містять модулі розпізнавання та порівняння текстового контенту веб-ресурсів із елементами машинного навчання. При викладанні дисципліни «Якість програмного забезпечення та тестування» у частині метрики та атрибути якості розглянуто методи оцінки ефективності функціонування інформаційних веб-ресурсів у формі часу виконання запиту залежно від структурованості контенту, який дає можливість врахувати характерні для нелінійних систем особливості, такі як залежність збурень від поточних значень показника, циклічність процесів, стрибкоподібні зміни характеристики.

Декан факультету комп'ютерних інформаційних технологій, д.т.н., професор — Микола ДИВАК

Завідувач кафедри комп'ютерних наук к.т.н., доцент — Андрій ПУКАС

Доцент кафедри комп'ютерних наук, к.т.н., доцент — Михайло ШПІНТАЛЬ

ЗУНУ
№ 126-20/1687 від 06.11.2020

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Західноукраїнського
національного університету
д.е.н., проф. Зеновій ЗАДОРОЖНИЙ

«__» 27.10 2020 р.

АКТ

про використання результатів дисертаційної роботи
Ковбасістого Андрія Вікторовича

«Математичне та програмне забезпечення підтримки функціонування інформаційних веб-ресурсів за умов різної їх структурованості»

Комісія у складі: голови — декана факультету комп'ютерних інформаційних технологій, д.т.н., проф. Дивака М.П. та членів: начальника науково-дослідної частини Письменного В.І., начальника планово-фінансового відділу Кушніра О.Р. склали цей акт про те, що дослідження та результати дисертаційної роботи Ковбасістого А. В. використані під час виконання науково-дослідних робіт на кафедрі комп'ютерних наук факультету комп'ютерних інформаційних технологій з безпосередньою участю автора, а саме:

- держбюджетної науково-технічної (експериментальної) розробки молодих вчених на тему: «Математичне та програмне забезпечення для контролю забруднення атмосфери автотранспортом» (державний реєстраційний номер 0116U005507);
- держбюджетного прикладного дослідження на тему: «Математичне та програмне забезпечення для класифікації тканин хірургічної рани в процесі операції на органах шиї» (державний реєстраційний номер 0117U000410);
- держбюджетного прикладного дослідження на тему: «Математичне та програмне забезпечення для ідентифікації та моніторингу особливо небезпечних джерел забруднення ґрунту та ґрунтових вод» (державний реєстраційний номер 0120U102040);
- госпдоговірної науково-дослідної роботи на тему «Онлайн система “Терногаз”» (державний реєстраційний номер 0119U102841);
- науково-дослідної роботи на тему: «Математичне та програмне забезпечення складних систем в умовах структурної та параметричної

невизначеностей» (державний реєстраційний номер 0117U000145); у яких автором розроблено математичне та програмне забезпечення для автоматизованого структурування та розпізнавання контенту із використанням засобів машинного навчання, яке у сукупності забезпечує підвищення ефективності функціонування інформаційних веб-ресурсів.

Голова комісії
декан факультету комп'ютерних
інформаційних технологій,
д.т.н., проф. Микола ДИВАК

Члени комісії:
начальник науково-дослідної
частини Валерій ПИСЬМЕННИЙ

Начальник
планово-фінансового відділу Олексій КУШНІР