УДК 004.75:004.65

В.І. КУНЧЕНКО-ХАРЧЕНКО
Черкаський державний технологічний університет

МЕХАНІЗМ ЛІНГВІСТИЧНОГО АНАЛІЗУ ЗБЕРЕЖЕНИХ ДАНИХ WWW-РЕСУРСУ ДЛЯ ПРОГНОЗУВАННЯ ЦИКЛІВ РОЗВИТКУ ІНТЕГРОВАНИХ СИСТЕМ

З метою формування механізму лінгвістичного аналізу збережених даних www-ресурсу для прогнозування циклів розвитку інтегрованих систем було проаналізовано статистичні показники інтегрованих інформаційних систем. В результаті чого виявлено хвилеподібний порядок розподілу збережених даних www-ресурсу для формування вектору часового ряду. Запропонований механізм реалізації наведеного теоретичного апарату у вигляді семантичної мережі лінгвістичного процесору, що узгоджується з концепцією об'єктно-орієнтованого програмування.

Ключові слова: семантична мережа, лінгвістичний аналіз, прогнозування, інтегрована система, часовий ряд, об'єктно-орієнтоване програмування.

V.I. KUNCHENKO-KHARCHENKO
Cherkassky State Technological University

MECHANISM OF LINGUISTIC ANALYZE SAVED DATA OF WWW RESOURCE FOR THE FORECASTING OF DEVELOPMENTS' CYCLES OF THE INTEGRATED SYSTEMS

Abstract - Was analyzed statistical volumes of the integrated systems for the forming of the mechanism of linguistic analyze of www-resources' saved data for the forecasting of developments' cycles of the integrated systems. As the result was defined the wave-formed order of the distribution of the saved www-resource data. These ones are used for the forming of time series' vector. The proposed mechanism of the theory as the semantic net of the linguistic processor is an accordance of the objective programming conception.

Key words: semantic net, linguistic analyze, forecasting, integrated system, time series, objective programming.

Вступ

На сучасному етапі розвитку суспільства джерела інформації можуть мати різнорідний характер. Використання інформації для побудови точних прогнозів виправдовує отримання рідкісної та достовірної інформації. В практиці часто прагнуть знайти деякі особливі джерела такої інформації, в т.ч. і нелегальні.

Аналіз досліджень за тематикою. З даними джерел [1, 2] біля 90% потрібної, існуючої інформації для прогнозу можна використати із легальних джерел. Різнорідний характер такої інформації наводиться на рис. 1. Потенційно цінним джерелом такої інформації можуть виступати збережені дані www-ресурсу. Для створення прогнозів циклів розвитку інформація повинна бути ефективно відфільтрована, відсортована, та перетворена в часовий ряд визначеного типу вектора чи векторного поля [3].

Формулювання предметної області. На даний момент не існує чіткої концепції експертної системи, алгоритму, що дає змогу отримувати збережені дані www-ресурсів (cookie) для формування часових рядів для систем прогнозування. Складнощі у виборі джерела інформації для процесу планування також виникають через велику їх велику кількість (рис. 1).

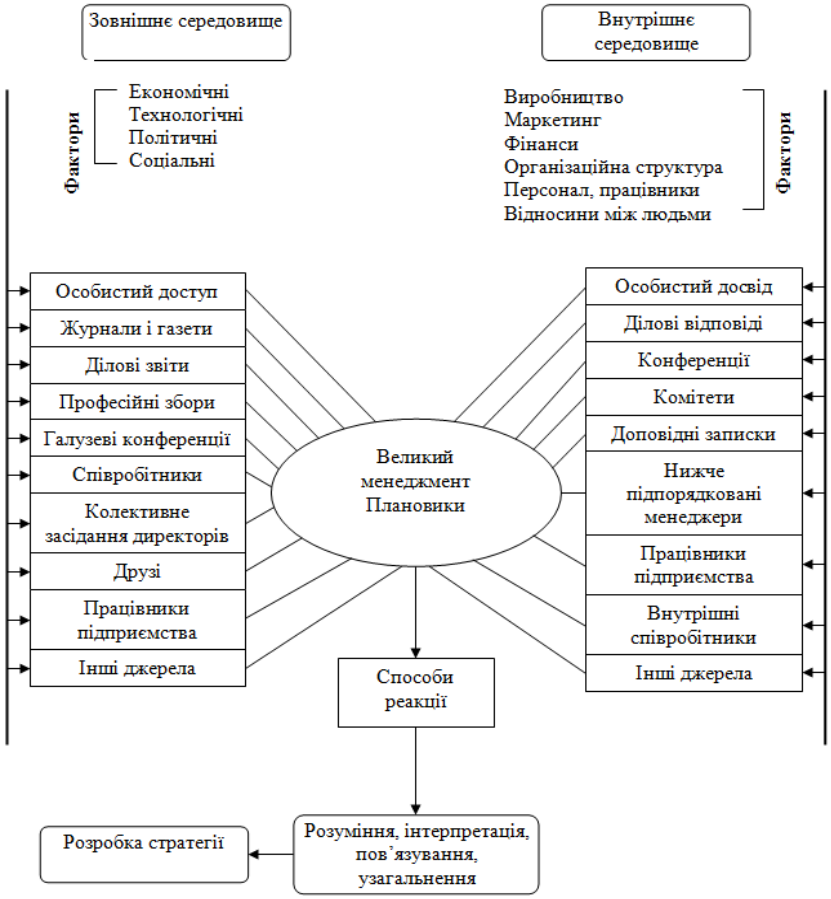


Рис. 1. Джерела інформації для процесу планування

Виклад основного матеріалу. Для з'ясування потенційної практичної цінності та вибору методу формування часового ряду для прогнозування визначимо проаналізуємо статистичні показники даних візитів на сайт інтернет-магазину, отриману через службу Google Analytics [4]. Вибрані для опрацювання дані з даної служби наведені в таблиці 1.

Таблиця 1

Статистичні показники даних візитів на сайт інтернет-магазину Google Analytics

Місяць	Унікальні відвідувачі	Кількість візитів сумарна	Різниця	Повна кількість перегляду сторінок ресурсу	Обсяг спожитого трафіку, Гб
1	2	3	4	5	6
Січень 2010	49260	64900	15640	4929738	29,14
Лютий 2010	102609	150857	48248	11341949	71,39
Березень 2010	120732	181087	60355	13774720	88,87
Квітень 2010	148756	226615	77859	17538939	106,98
Травень 2010	175433	264182	88749	20709849	122,56
Червень 2010	202611	309055	106444	24015381	141,64
Липень 2010	196101	291979	95878	22628495	134,76
Серпень 2010	146925	215512	68587	16558195	99,97
Вересень 2010	110517	161719	51202	12471780	75,69
Жовтень 2010	110487	161767	51280	12149094	74,69
Листопад 2010	134764	198883	64119	15657377	104,43
Грудень 2010	133989	195775	61786	15175711	97,85

В результаті побудови графічних залежностей розподілів унікальних відвідувачів (рис. 2) та сумарної кількості відвідувачів в часі (рис. 3) можна зробити висновок про хвилеподібний характер даних явищ.

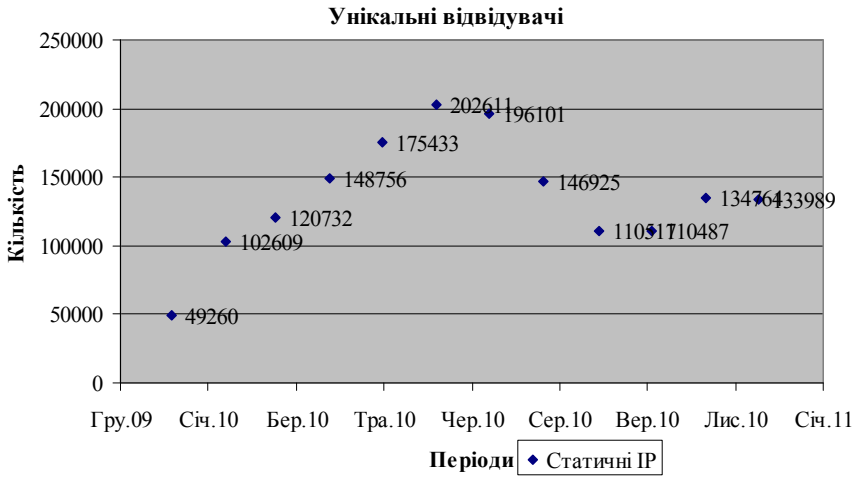


Рис. 2. Графічна залежність розподілів унікальних відвідувачів в часі

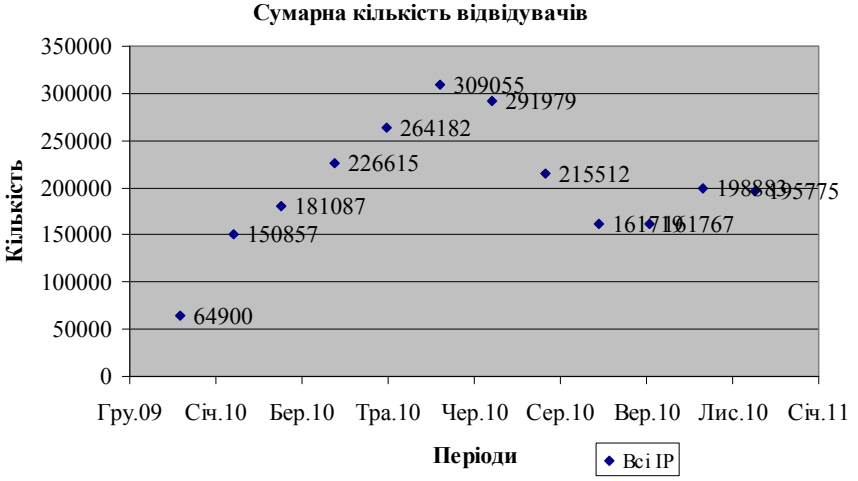


Рис. 3. Кількість візитів сумарна

Можна припустити, що відповідні провали на графіках (рис. 2, 3) є сезонними провалами попиту на

графіках у вихідні, святкові дні. А порівнюючи різницеві дані по місяцям на рис. 4 можна вести мову про можливість виводу корегуючої функції для часового ряду, подібної до наведеної на рис. 4.

Якщо говорити про статистику походження запитів до сайту інтернет-магазину, отриману через Google Analytics [4], то з великою часткою ймовірності загальна кількість відвідувачів сайту перевищує кількість користувачів з унікальними IP-адресами через використання ними протоколу DHCP, відповідно до якої надається IP-адреса для доступу до глобальної мережі робочим станціям на рівні протоколу TCP/IP.

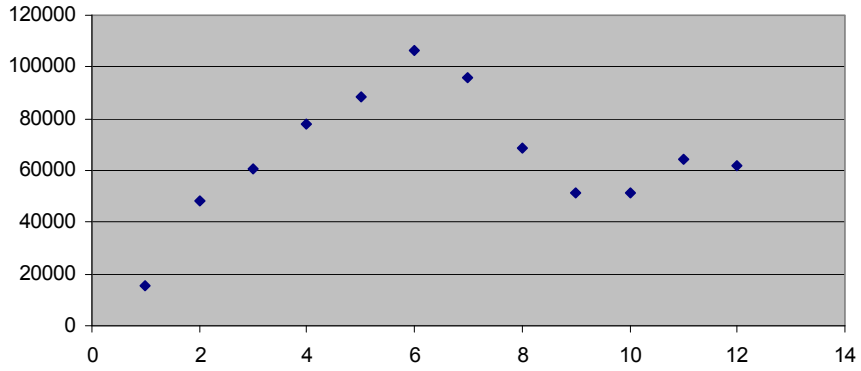
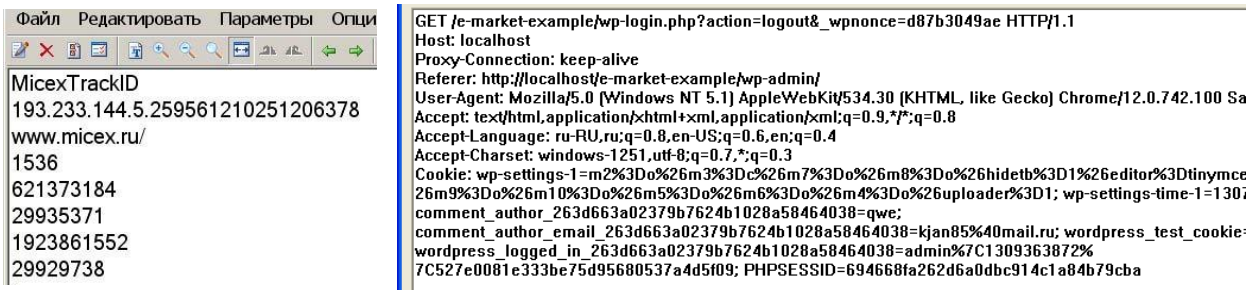


Рис. 4. Різницеві дані кількості відвідувачів

Технологія DHCP дозволяє надавати IP-адреси великій кількості клієнтів в межах декади простору масок адресації сегменту мережі. Верхня і нижня її можливі межі мають значення $\sup IP = XXX.XXX.XXX.255$ і $\inf IP = XXX.XXX.XXX.0$ відповідно [5]. Що може означати той факт, що деякі з користувачів відвідують інтернет-магазин, заходячи з великих сегментів мереж. По точкам графіку рис. 4 можна відновити функціональну залежність, наприклад, за методом Кунченко [6].

Як видно з викладеного вище, використання даних служб Google Analytics дає доволі об'єктивний часовий ряд для подальшого прогнозування, але використання таких служб не завжди доступно, користування службами Google Analytics особливо в розділі комерційної інформації може бути пов'язано зі значними витратами і неможливістю бачити всі отримані дані та деякими іншими причинами. Розв'язок проблеми може критися у використанні збережених даних www-ресурсу (Cookie).

Фактично, cookie – це невеликий обсяг текстової інформації (типу String), яку сервер передає браузеру. Браузер буде зберігати цю інформацію і передавати її серверу з кожним запитом як частину HTTP заголовка. Одні значення cookie можуть зберігатися тільки протягом однієї сесії, вони видаляються після закриття браузера. Інші cookie-файли встановлені на деякий період часу, записуються у файл. Цю властивість механізму роботи з cookie необхідно врахувати при формуванні часового ряду. Зазвичай файл з такою інформацією називається 'cookies.txt' і лежить в робочій директорії встановленого на комп'ютері браузера. Приклад вмісту такого файлу в браузері і сервері наведено на рис. 5а і 5б відповідно.



а **б**

Рис. 5. Приклад вмісту cookie-файлу

Як видно з рис. 5а, 5б, зміст cookie-файлу містить інформацію, яка потенційно може використовуватися для формування часових рядів різного типу даних (String, Integer тощо). По аналогії з традиційними документами, файл збережених даних www-ресурсу для прогнозування циклів розвитку інтегрованих систем може бути розкладений на семантичні елементи у відповідності до кодів реквізитів згідно з уніфікованою системою організаційно-розпорядчих документів (УСОРД) [7], а далі оброблений за допомогою ЛП. Підготовка даних, збережених в такому файлі вимагає створення класу ЛП, що орієнтовані на обробку текстів характерної мови, що зберігаються в базі cookie-файлів, а саме: інформаційних блоків IP-адрес; URL, URI посилань; заголовків та ключів форматованих web-ресурсів тощо.

Семантична мережа ЛП формується на основі обробленої інформації структуру бази знань (БЗ). ЛП виділяє з текстів семантично значущу складову. У текстах cookie-файлів ЛП повинен виділяти тільки ту інформацію, яка необхідна для побудови прогнозу і подальшої автоматичної побудови рішень. Наприклад, це можуть бути конкретні посадові особи, назви організацій, якісні характеристики товарів, послуг, локація, ціна.

Як зазначено в [8], ЛП забезпечує автоматичну побудову змістовних образів. Він включає в себе елементи лексикографічного, морфологічного, термінологічного і семантико-синтаксичного аналізу та апарат формування алфавіту.

Блок лексикографічного аналізу ЛП забезпечує: 1) автоматичний поділ тексту на самостійні частини (наприклад, виділення документів із загального обсягу інформаційного потоку); 2) визначення початку та кінця тексту; 3) автоматичне визначення цифрових блоків та блоків службових символів у тексті.

При морфологічному аналізі кожному слову надаються ознаки, які діляться на три групи 1) лексичні (слово з великої літери, з точкою на кінці або це окрема буква та ін.); 2) морфологічні (граматична категорія слова: рід, число, відмінок, відміна для іменника); 3) семантичні (прізвище, ім'я та по батькові) [9, 10].

Кількість семантичних ознак може збільшитися як за рахунок спеціальних словників – словників полів кодів реквізитів документів, специфічних кваліфікаційних ознак тощо [10]. Саме слово в нормальній формі також вважається ознакою. Блок морфологічного аналізу заснований на узагальнених закінченнях слів Результатом роботи блоку морфологічного аналізу є семантична мережа, що є ієрархічною просторовою структурою тексту документа (рис. 6) та оброблює початковий текст (алфавіт).

У такій структурі (рис. 6) наведені слова в нормальній формі з їх ознаками та вказуванням їх послідовності. Подальша обробка зводиться до перетворення мереж на основі заданих правил. Для виділення необхідного інформаційного векторного потоку достатньо знайти в семантичній мережі для ЛП цифрові або текстові результати, що відповідають шуканому параметру часового ряду, передати їх експертній системі для з'ясовування рівня пошукового шуму, знаходження рівня оптимальності критеріям релевантності. Цьому передує термінологічний аналіз.

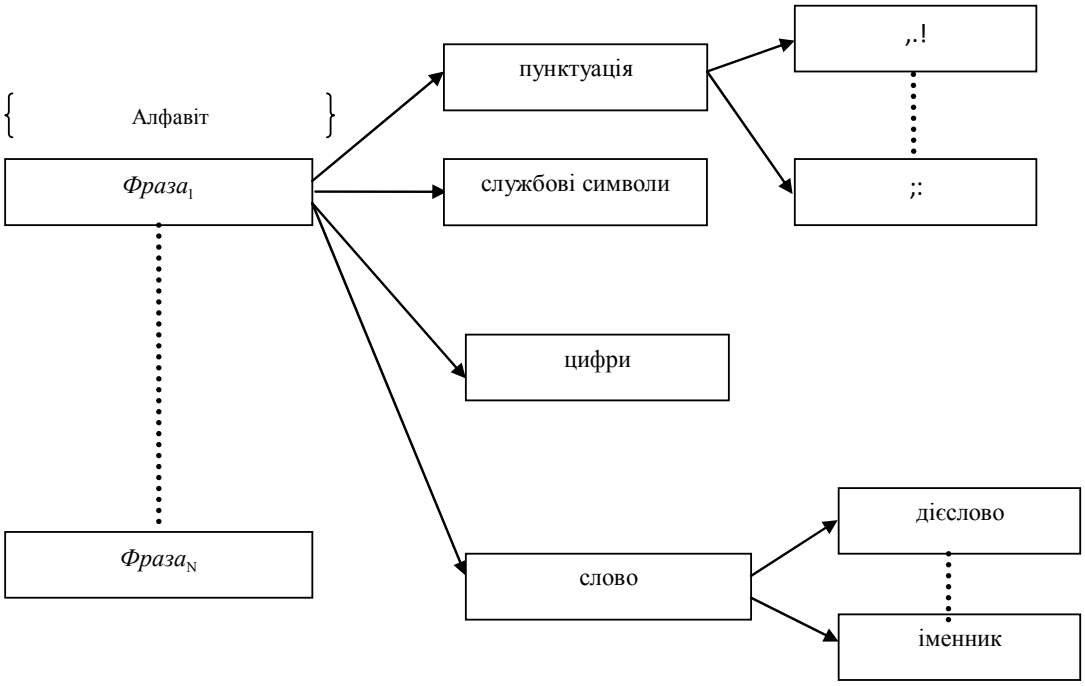


Рис. 6. Семантична мережа для ЛП

Синтаксичний аналізатор синтезується на базі кінцевого автомату (КА). В даному випадку використовується уявлення мови, що аналізується у вигляді деякого виду КА .

Реалізація викладеного в статті механізму, здійснена на базі об'єктно-орієнтованої ООП J2SE 7.0.

В якості частини реалізації обробника потоку даних, використано J2SE CookieHandler з набором відповідних методів. Даний клас дозволяє керувати станом системи за допомогою cookie. По закінчення сесії з'єднання cookie-файли передаються до БД. Де зберігаються до моменту початку опрацювання їх ЛП.

Для роботи з ЛП засобами ООП Java створено новий клас StringLP, наведений нижче:

```
public class StringLP {
  public static TCDB (String str, int a, String sa, int b, String sb, String
sc, int c, String sd, int d ){
    //------Код методу ===============
  }
}
```

Для нього об'явлено статичний метод TCBD в якому наявні девять аргументів, які відповідають чотирьом блокам виділення морфем семантичною мережею (рис.6):

1) str – вихідний рядок;

2) a – номер входження першого підрядка;
3) sa – перший підрядок;
4) b – індекс входження другого підрядка;
5) sb – другий підрядок;
6) c – індекс входження третього підрядка;
7) sc – третій підрядок;
8) d – індекс входження четвертого підрядка;
9) sd – четвертий підрядок.

Клас String в Java призначений для зберігання рядків. Створити екземпляр класу можливо простим присвоюванням: String str=”string”. Даний метод є статичним, для його виклику не потрібно створювати екземпляр класу. Змінна result – кінцевий результат, який повертає клас StringLP.

Висновки. В результаті проведеної роботи з формування механізму лінгвістичного аналізу збережених даних www-ресурсу для прогнозування циклів розвитку інтегрованих систем проаналізовано статистичні показники інформаційних систем та з’ясовано хвилеподібний порядок розподілу статистичних даних для формування вектору часового ряду. Створена семантична мережа дозволяє виділяти специфічний алфавіт з бази знань збережених даних www-ресурсу. В результаті, механізм реалізації наведеного теоретичного апарату здійснено в рамках концепції ООП Java. В рамках подальших досліджень потребує розв’язку задача відновлення аналітичного виразу функціональної залежності різницевих даних (рис. 4) та побудова експертної системи. На базі методу визначення схожості повних текстів статистичної міри tf-idf, описаному в [15] може бути здійснена онтологічна формалізація отриманих часових рядів та їх сортування для підвищення точності прогнозів.

Література

1. Кунченко-Харченко В.І. Інформаційно-системні технології, архівознавство і документологія для прогнозу циклів розвитку соціальних та виробничих систем / В.І. Кунченко-Харченко. – Львів : УАД, 2009. – 300 с.

2. Макаров А.М. Маркетинг : учебное пособие / А.М. Макаров. – Ижевск : Изд-во Института экономики и управления УдГУ, 2000. – 222 с.

3. Abderrahman Atmani. A fuzzy expert system for automatic seismic signal classification / Abderrahman Atmani, El Hassan Ait Laasri, Es-Said Akhouayri, Dris Agliz, Daniele Zonta // Expert Systems with Applications. Volume 42, Issue 3, 15 February 2015, Pages 1013–1027.

4. Анализ работы сайта Интернет-магазина с помощью Google Analytics [Електронний ресурс]. – Режим доступу : https://www.emagazin.info/ru/google_analytics_shop - 30.11.2015.

5. DHCP Primer [Електронний ресурс]. – Режим доступу : http://www.esubnet.com/fragment-dhcp-primer.html. – 01.11.2015.

6. Кунченко Ю.П. Поліноми наближення у просторі з породжувальним елементом / Ю.П. Кунченко ; пер. з рос. – К. : Наукова думка, 2005. – 219 с.

7. Державна уніфікована система документації. Уніфікована система організаційно-розпорядчої документації. Вимоги до оформлення документів. ДСТУ 4163-2003 [Електронний ресурс]. – Режим доступу : http://www.gerelo.dp.ua/index/info_dstu_4163-2003.html.

8. Сікора Л. С. Логіко когнітивна структура операційних стадій розв’язання задач управління ПНО / Л. С. Сікора, Р. Л. Ткачук, Т. Є. Рак, В. І. Кунченко-Харченко // Зб. наук. праць ІПМЕ ім. Г.Є. Пухова НАН України. – К., 2013. – Вип. 67. – С. 148–157.

9. Кузнецов И.П. Семантико-ориентированные системы на основе баз знаний : монография / И.П. Кузнецов, А.Г. Мацкевич. – М. : МТУСИ, 2007. – 173 с.

10. Кузнецов И.П. Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данних / И.П. Кузнецов, А.Г. Мацкевич // Труды международной конференции "Диалог 2006". – Бекасово, 2006. – С. 317–322.

11. Hlava M. Multilingual machine indexing / Hlava M., Hainebach R. // Proceedings of the Ninth international conference on new information technology, Pretoria, South Africa, november 11–14, 1996. – P. 105–120.

12. Beuster G. MIC – A System for Classification of Structured and Unstructured Texts. Diploma Thesis. – University Koblenz, 2001.

13. Berners-Lee T. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential / T. Berners-Lee. The MI Press, 2005.

14. Steinberger R. using thesauri for automatic indexing and visualisation / Steinberger R., Hagman J., Scheer St. // OntoLex, 2000. – 2000. – pp. 130–141.

15. Гранік М.О. Метод визначення схожості новинних текстів на основі статистичної міри «term-inverce document frequency» / М.О. Гранік, В.І. Месюра // Вісник Хмельницького національного університету. – 2015. – № 4. – С. 180–182.

Рецензія/Peer review : 21.11.2015 р. Надрукована/Printed :6.12.2015 р.
Рецензент: д.т.н., проф. Лега Ю.Г.