УДК 81.33

Надутенко М. В.,
кандидат технічних наук,
завідувач відділу інформаційних технологій
Українського мовно-інформаційного фонду
Національної академії наук України

Старишко Ю. В.,
молодший науковий співробітник
Українського мовно-інформаційного фонду
Національної академії наук України

СИСТЕМА ПРИРОДНОМОВНОГО АНАЛІЗУ КОРПУСНОГО ТИПУ ЯК ЗАСІБ ОПРАЦЮВАННЯ ІНТЕРНЕТ-ВИДАНЬ

Анотація. У статті проаналізовано сучасний світовий обсяг створених і реплікованих людством даних. На основі цього аналізу розкрито проблематику збору та аналізу даних в Інтернет-мережі загалом та в інтернет-ЗМІ зокрема. Розроблено Систему природномовного аналізу корпусного типу, що адаптована до великих масивів різноформатних багатомовних даних. Проілюстровано застосування цієї системи на прикладі науково-освітнього веб-порталу.

Ключові слова: система природномовного аналізу, корпус текстів, автоматичний аналіз інтернет-видань, краулінг інтернет-сторінок.

Постановка проблеми. За даними організації IDC, у 2016 р. загальний світовий обсяг створених і реплікованих людством даних становив більше 16 зеттабайт (ЗБ) (16 трильйонів ГБ). За прогнозами цієї ж організації до 2025 р. обсяг даних на планеті збільшиться в 10 разів і становитиме 163 ЗБ. (Для порівняння, увесь світовий обсяг інтернет-трафіку в 2016 р. перевищив 1 ЗБ, а ще в 2006 р. обсяг інформації, вироблений людством за всю свою історію, дорівнював 0,16 ЗБ). Попри безпрецедентне зростання інформації, у світі, за оцінками IDC, тільки 0,4% даних аналізується. Закономірно постає проблема опрацювання великих масивів даних, що зумовлює необхідність створення інтелектуальних систем опрацювання природномовної інформації. Індивідуальна пошукова система (адаптована до великих масивів даних, націлена на різноформатність, багатомовність) забезпечить широкі можливості індексації корпусів текстів.

Аналіз останніх досліджень і публікацій. На сьогодні пошукових систем існує досить багато і кожна виконує свої завдання (часто – комерційні). Попри це, створення індивідуальної системи інформаційного пошуку не втрачає актуальності, оскільки це дає змогу формувати власні великі індексовані масиви текстів з гіперпосиланнями. Роботу пошукової системи умовно можна поділити на три безперервні процеси: краулінг інтернет-сторінок, індексація отриманих даних і ранжування. Цій проблематиці приділяли увагу іноземні (Pant G., Srinivasan P., Menczer F., Aggarwal C.C., Menczer F., Pant G., Srinivasan P. Micarelli A., Gasparetti F., Castillo C., Marin M., Rodriguez A. Ester M., Kriegel H.P., Pant G., Menczer F., Hesham A., Liu H., Milios E., Janssen J., Rennie J., McCallum A.K., Diligenti M., De Bra P., Rungsawang A., Angkawattanawit N., Aggarwal C.C., тощо) та українські (Замятін Д.С., Михайлюк А.Ю., Михайлюк В.А., Петрашенко А.В., Циганкова К.Р., Шаповаленко Є.І., Широков В.А., Шевченко І.В., Рабулець О.Г. тощо) дослідники, але, як уже зазначалося, створення індивідуальної системи інформаційного пошуку, що включає інтернет-краулінг та автоматичне індексування отриманого тексту, є важливим і актуальним завданням, оскільки це стане основним джерелом наповнення корпусу текстів.

Мета статті – описати функціонал, технічні та користувацькі можливості Системи природномовного аналізу корпусного типу; проілюструвати застосування цієї системи на прикладі науково-освітнього веб-порталу для аналізу інтернет-видань.

Виклад основного матеріалу. Описаний нами стан сучасного інформаційного інтернет-середовища зумовлює необхідність створення систем, які без перешкод на великих масивах працюють із різноформатними та багатомовними файлами й адаптовані до подальшого аналізу природномовних об'єктів. Одна з таких систем розроблена співробітниками Українського мовно-інформаційного фонду Національної академії наук України спільно з Інститутом телекомунікацій та глобального інформаційного простору НАН України.

До створеної системи підключена підсистема краулінгу, яка є тісно інтегрованою з корпусною системою, системою індексації та полімовною синонімічною зоною. Краулери, як і корпусна система, є віртуалізованими лексикографічними агентами, тобто є різновидом лексикографічних систем. Ця система є першим кроком наповнення корпусу текстів і дозволяє в рази мінімізувати використання людських ресурсів у цьому напрямку роботи.

Вхідні дані подаються до системи у вигляді початкового списку сайтів та списку пошукових запитів. Вихідними даними системи є списки релевантних фрагментів тексту з активними посиланнями на джерело.

Застосування цієї системи проілюстровано на прикладі науково-освітнього веб-порталу «Тарас Шевченко».

Пошуковий апарат розглянутої системи дозволяє обрати індивідуальні параметри краулера та індексатора із зазначенням максимальної глибини краулінгу сторінок (див. Рис. 1. – Налаштування параметрів краулера та індексатора).

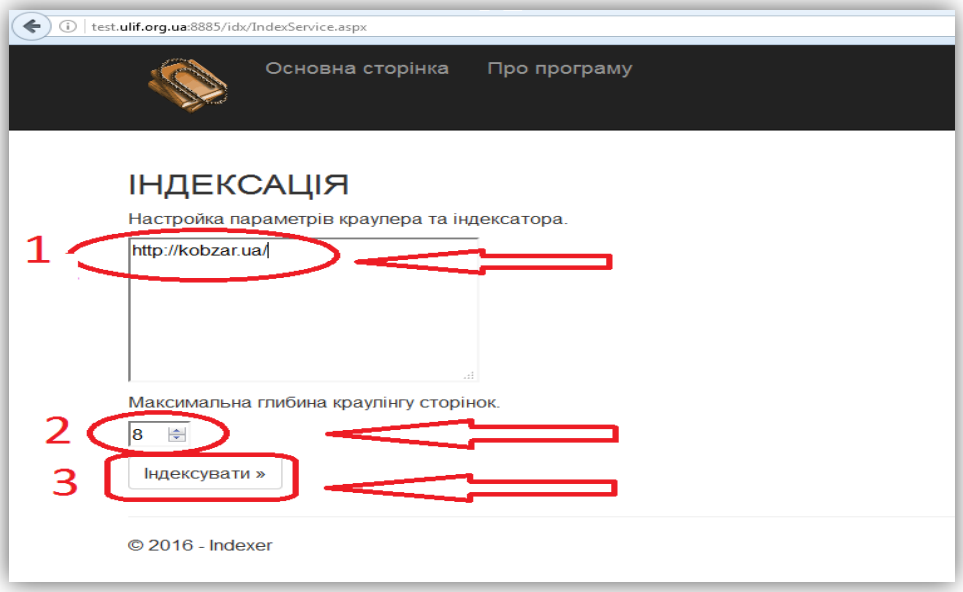

Рис. 1. Налаштування параметрів краулера та індексатора

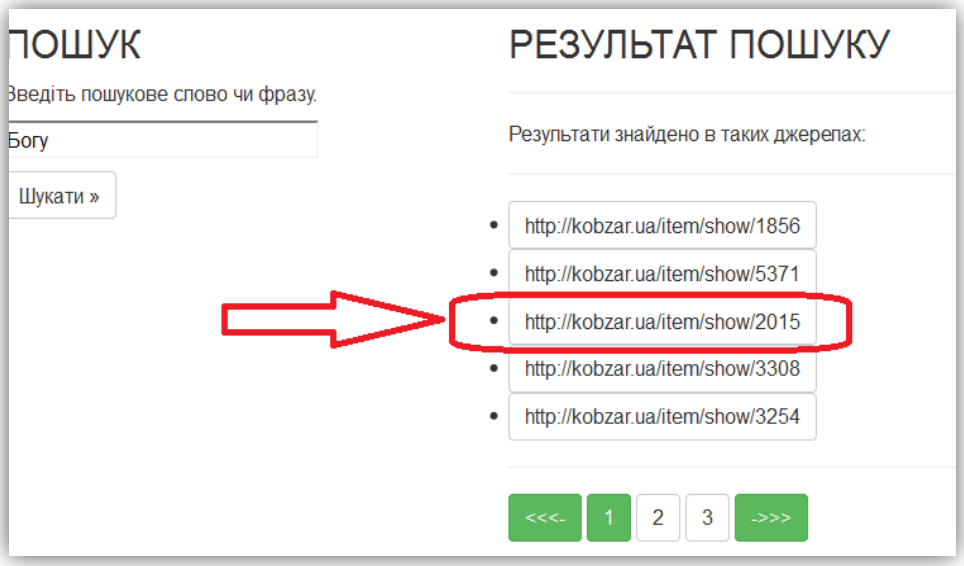

Рис. 2. Результат пошуку

Наступним діалоговим вікном є завдання пошукового слова. Результатом повнотекстового пошуку є список посилань (Рис. 2. Результат пошуку) з доступом до кожної локації пошукового фрагмента в тексті, тобто до всіх контекстів, які містить пошуковий фрагмент, із функцією відкрити посилання джерела контексту (Рис. 3. Результати повнотекстового пошуку).

Застосування спеціалізованих технологій онтологокерованого веб- чи інтернет-краулінгу дозволяє в подальшому створити систематизовану колекцію мережевих текстів, об'єднаних за однією або сукупністю ознак (мовних, понятійних, прагматичних, часових, стильових, функціональних, інтенціональних тощо). Найбільш затребуваними є колекції текстів однієї тематики (навчальні та наукові колекції), одного автора (повне зібрання творів), певної історичної епохи, певною мовою (електронні бібліотеки Національних лінгвістичних корпусів) або створених за певних обставин, у певній формі, з певною метою (навчально-методичні матеріали, нормативно-правові акти, що регулюють правовідносини у визначеній сфері тощо), для категорії читачів із певним рівнем доступу (публічні дані, дані для службового користування) тощо. Колекція текстів може насамперед стати інструментом дослідження різних інтра- та екстрамовних фактів.

Запропоновані нами підходи до створення інформаційно-комунікаційних і корпусних технологій дозволять у подальшому робити це динамічно, вибираючи релевантні запиту користувача повнотекстові документи з вебу – супермасиву проіндексованих текстів або локальних баз – спеціалізованих електронних бібліотек. На рисунках 4–7 представлено систе-

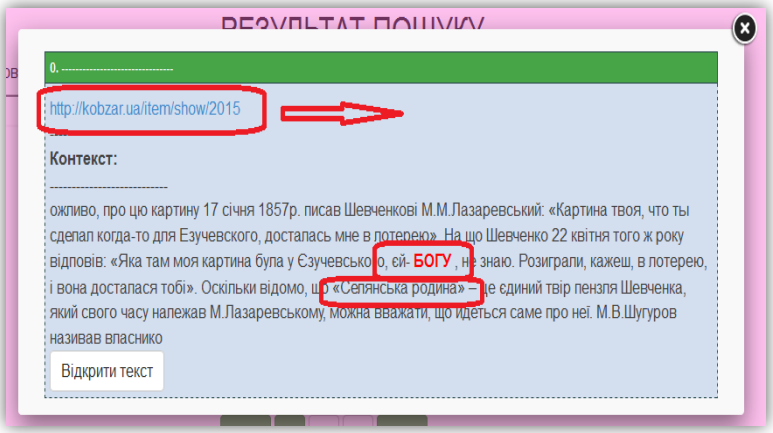

Рис. 3. Результати повнотекстового пошуку

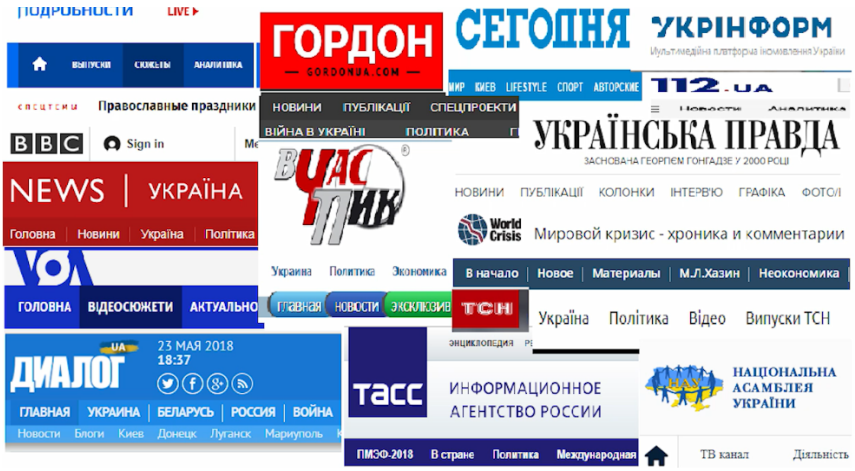

Рис. 4. Пошукова призма ТОДОС: джерела даних

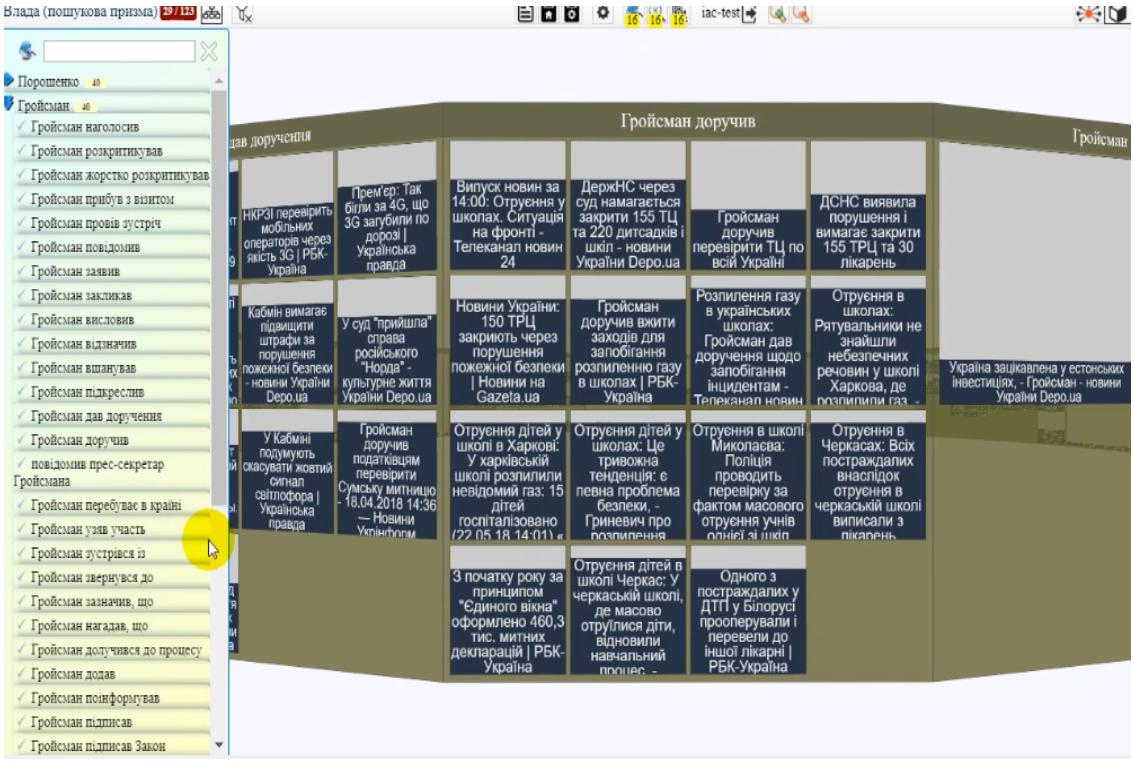

Рис. 5. ПРИЗМА новин за тематичним напрямом «влада»

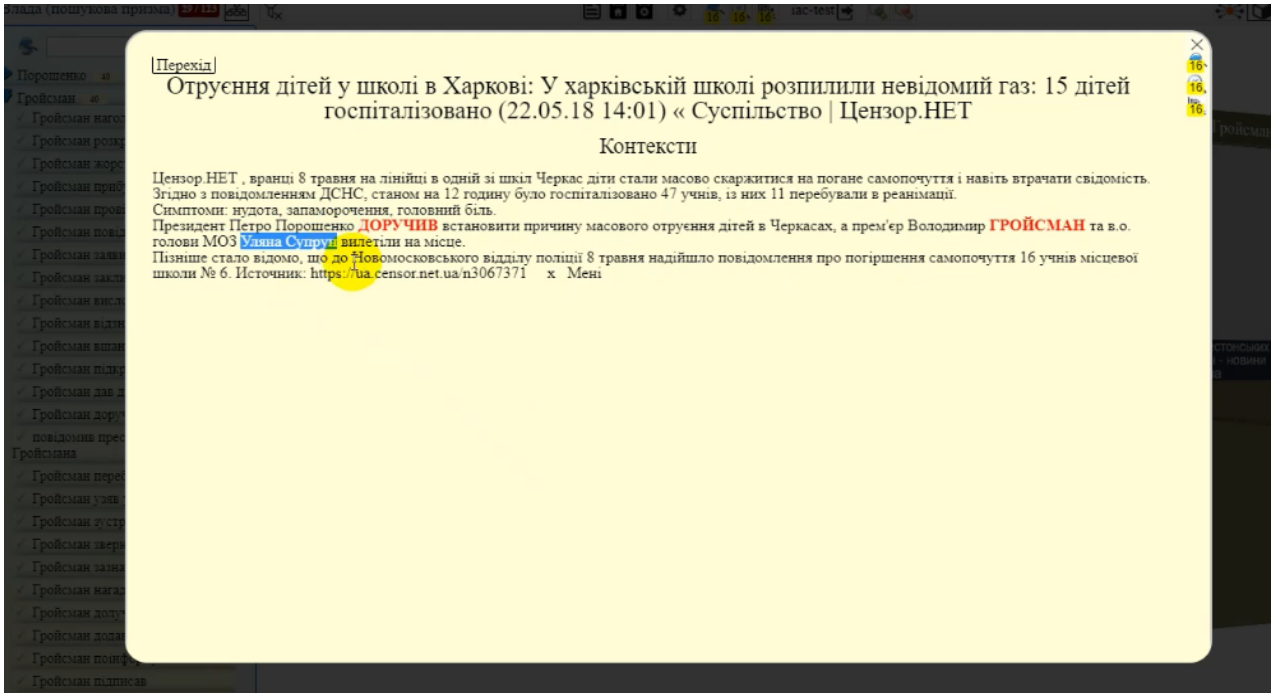

Рис. 6. Перегляд контекстів та пошук контекстно пов'язаних концептів

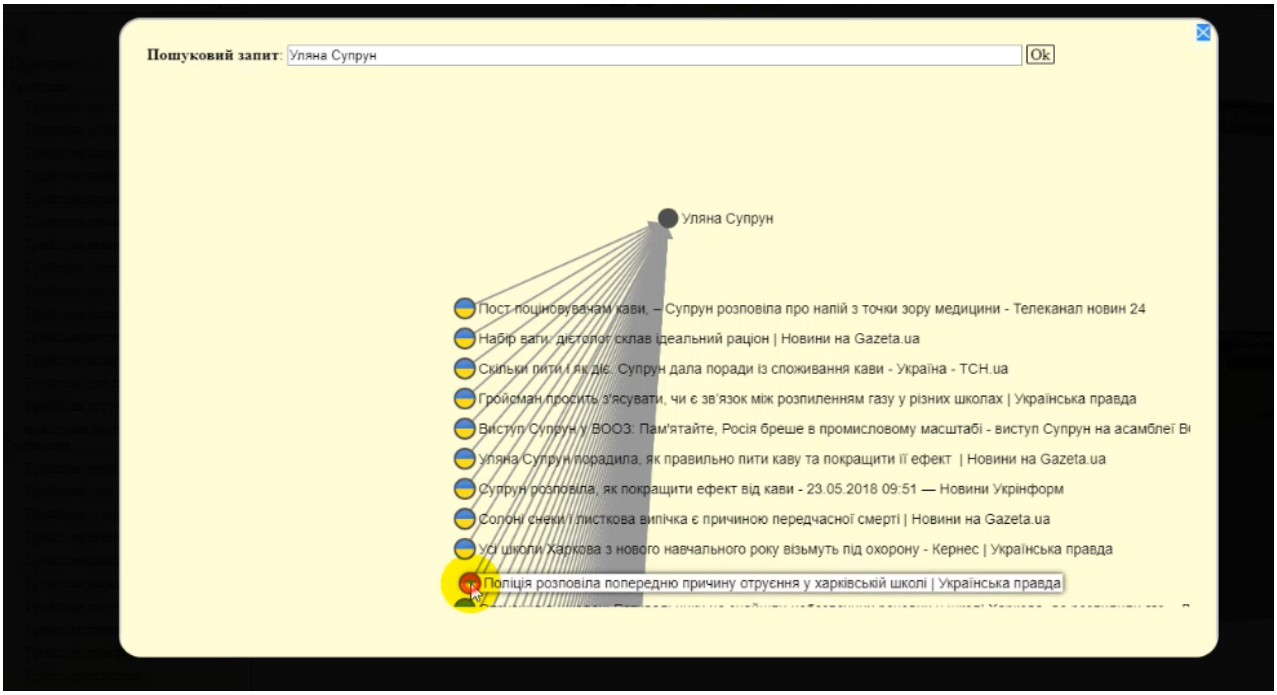

Рис. 7. Результат пошуку контексто пов'язаних концептів

му природномовного аналізу корпусного типу, яка є модулем ІТ-платформи ТОДОС (Трансдисциплінарні Онтологічні Діалоги Об'єктно-орієнтовних систем), з її допомогою здійснено пошук контекстно пов'язаних концептів ЗМІ.

Висновки. Таким чином, через швидке поширення інформації, особливості формату та відсутність географічних обмежень збір та аналіз даних в Інтернет-мережі загалом та в інтернет-ЗМІ зокрема потребують використання спеціальних технологій. Презентовані можливості новоствореної нами системи достатньо широкі. Використовуючи її під час моніторингу інтернет-видань, із мінімальними витратами часу та людських зусиль дослідник може отримати найповнішу інформацію за певним запитом.

Література:

1. Широков В.А, Шевченко І.В., Рабулець О.Г. Природномовна індексація як засіб вдосконалення пошукового апарату інформаційних систем. *НТІ.* 2000. № 3. С. 23–25.
2. Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы. *«Информационные технологии».* Москва : Вид. «Новые технологии», 2009. С. 50–55.

Надутенко М. В., Старишко Ю. В. Система корпусного типа анализа естественного языка как средство обработки интернет-изданий

Аннотация. В статье проанализирован современный мировой объем созданных и реплицированных человечеством данных. На основе анализа раскрыта проблематика сбора и анализа данных в Интернет-сети в целом и в интернет-СМИ в частности. Разработана Система анализа естественного языка корпусного типа, которая адаптирована к работе с большими разноформатными и многоязычными массивами данных и к дальнейшему анализу лингвистических объектов. Проиллюстрировано применение этой системы на примере научно-образовательного веб-портала.

Ключевые слова: система естественного анализа, корпус текстов, автоматический анализ интернет-изданий, краулинг интернет-страниц.

Nadutenko M., Staryshko Yu. The system of natural-language analysis of text corpora as a mean of processing of Internet publications

Summary. The article analyzes the modern world of data created and replicated by humanity. The problems of collecting and analyzing data in the Internet network in general and in Internet media in particular are disclosed. A system of natural-language analysis for the personal use, adapted for large multilingual and multiformat data sets and further analysis of natural-language objects was developed. The application of this system is illustrated on the example of the scientific and educational web-portal.

Key words: system of natural-language analysis, corpus of texts, automated analysis of Internet publications, crawling of Internet pages.