ISSN 2076-1546

УДК 621.39 DOI: 10.46972/2076-1546.2023.25.05

Ю. І. Міхєєв, В. В. Лобода, Т. М. Войтко, Р. І. Гладич

АВТОМАТИЗОВАНА СИСТЕМА ПОШУКУ ІНФОРМАЦІЇ В МЕРЕЖІ ІНТЕРНЕТ

У статті розглянуто процес пошуку інформації за визначеною тематикою в мережі Інтернет. Увагу акцентовано на особливостях, які пов'язані з функціонуванням інформаційно-пошукової системи, досліджено роботу її типового зразка. Така система має забезпечувати гнучку реакцію на запит оператора шляхом інтерактивного пошуку й підтримання архіву запитів, включаючи тезаурус, засоби перевірки орфографії та пунктуації мовного запиту. Результати проведеного аналізу наявних автоматизованих систем пошуку інформації свідчать про те, що запит має формуватися з урахуванням: використання логічних зв'язків, обмежень у відстані між словами, можливості встановлення вагових коефіцієнтів ключовим словам, сортування за датою та розміром документа.

Запропоновано підхід до створення автоматизованої системи пошуку тематичної інформації в Інтернеті, що базується на унікальному характері оперативно-аналітичної діяльності спеціальних підрозділів, а також її структурну схему. Така система має враховувати особливості роботи різноманітних інтернет-сервісів, які використовують аналітики спеціальних служб: тематичні каталоги ресурсів, сайти новин, RSS-повідомлення та інформаційні агентства, що транслюють новини онлайн. Передбачається, що розроблення спеціального програмного забезпечення на базі запропонованої функціональної структури автоматизованої системи пошуку тематичної інформації в мережі Інтернет дозволить підвищити ефективність інформаційно-аналітичної діяльності в спеціальних підрозділах за рахунок наявності засобів: пошуку різнорідної інформації за визначеними об'єктами; виявлення зв'язків між об'єктами моніторингу та пов'язаними з ними фактами й подіями; проведення візуалізації результатів аналітичних досліджень.

Ключові слова: *автоматизована система; мережа Інтернет; інформаційний запит; інформаційно-пошукова система; структурна схема.*

Постановка проблеми в загальному вигляді. Тенденції розвитку інформаційного простору пов'язані зі збільшенням масивів даних, які циркулюють у мережі Інтернет і відрізняються за типом, формою, тематикою та є здебільшого неструктурованими. Це призводить до ускладнення швидкого та релевантного пошуку потрібної інформації за допомогою стандартних інструментів інтернет-сервісів. У таких умовах необхідно передбачати використання різних методів збору інформації, які можуть використовуватися окремо або в комплексі разом зі спеціальними системами пошуку. Особливо це стосується інформаційно-аналітичної діяльності в інтересах спеціальних підрозділів. Для якісного забезпечення їх інформаційних потреб інформаційно-аналітична діяльність має передбачати виконання комплексу таких інформаційних процесів, як: пошук, оброблення та зберігання інформації.

Аналіз останніх досліджень і публікацій. На сьогоднішній день для пошуку інформації в мережі Інтернет використовують індивідуальні пошукові агенти, метапошукові системи, каталоги та системи інтернет-моніторингу [1, 2]. Аналіз результатів виконання конкретного завдання щодо тематичного пошуку інформації показав, що використання зазначених засобів дещо ускладнене через їх вузьку спрямованість. У наукових працях розглядаються різні підходи до обробки текстової інформації, алгоритми формування пошукового розпорядження тощо [2–4]. Однак системний пошук інформації за визначеною тематикою не досліджувався. Це призводить до неможливості ефективного використання запропонованих підходів в інформаційно-аналітичній діяльності в умовах забезпечення максимальної повноти інформаційного масиву, автоматизації процесу збору даних та пошуку засобів навігації в ній із подальшим вилученням необхідних знань.

Формулювання завдання дослідження. На сьогоднішній день потребує вирішення ціла низка питань, які стосуються функціонування інформаційно-пошукових систем. Головними з них є: розробка алгоритмів і способів раціонального пошуку й подальшої обробки даних за визначеною тематикою та, як наслідок, створення на їх основі автоматизованої системи пошуку інформації (АСПІ) за запитом у мережі Інтернет. Використання розробленої системи для виконання завдань інформаційно-аналітичної діяльності дозволить скоротити час, необхідний для підготовки звітів, що забезпечить більшу ефективність пошуку та аналізу даних спеціальними підрозділами. Отже, метою статті є розгляд питання з розроблення АСПІ в мережі Інтернет.

Виклад основного матеріалу. Будь-який пошук інформації починається з визначення інформаційної потреби, після чого оператор повинен сформувати пошуковий запит (рис. 1).

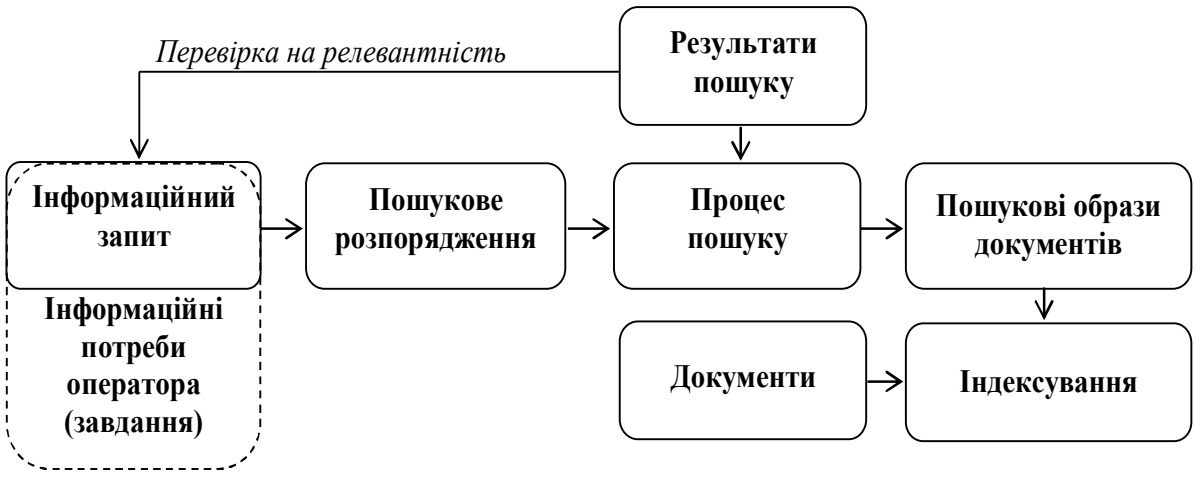

Рис. 1. Схема функціонування інформаційно-пошукової системи

Для виконання пошукового запиту, по-перше, необхідно скласти його пошуковий образ, тобто його формальне подання в термінах інформаційно-пошукової мови. Після цього складається відповідне розпорядження, яке містить у собі пошуковий образ запиту та вказівки на логічні операції, які виконуються. Далі воно порівнюється із пошуковими образами документів, у результаті отримуються відомості про адресу зберігання та

короткий зміст документа. Від того, наскільки повно буде з'ясовано завдання оператором, у подальшому залежатиме ступінь релевантності (відповідності) знайденого матеріалу змісту завдання [5].

Такий підхід до пошуку інформації відповідає векторній (лінійній) моделі, яка сьогодні застосовується в більшості відомих пошукових машин завдяки своїй простоті [4]. Для зображення цієї моделі у формалізованому вигляді використовують такі поняття, як: словник – упорядкована множина термінів потужності D; документ (пошуковий образ документа) – двійковий вектор розмірності D; база – матриця $L = N \times D$, рядки якої відповідають N документам. Якщо термін входить у документ, то у відповідному розряді вектора стоїть 1, інакше – 0. У цьому разі процедура обробки запиту набуває такого вигляду: $Lq = r$, де q – вектор запиту, а r – відгук системи на нього. Модель ускладнюється шляхом використання вагових коефіцієнтів у термінах документа та запиту, які відображають їх значущість. Використання вагових коефіцієнтів дозволяє визначити міру близькості "документ – запит" s як косинус кута між векторами запиту та документа:

$$s(q,d) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2 \sum_i d_i^2}}, \qquad (1)$$

де $0 \le s(q,d) \le 1$;

$q = (q_1,...,q_n)$ – запит, q_i – ваговий коефіцієнт i-го терміна запиту;

$d = (d_1,...,d_n)$ – документ, d_i – ваговий коефіцієнт i-го терміна документа.

Більш гнучку реакцію на запит оператора можна забезпечити шляхом інтерактивного пошуку. Суть такої моделі полягає у зворотному зв'язку за релевантністю. Оператору надаються початкові результати обробки знайденого, які він оцінює на відповідність. Після чого запит корегується та пошук продовжується. Крім того, АСПІ повинна підтримувати архів запитів, включати тезаурус, засоби перевірки орфографії та пунктуації. Для підвищення повноти пошуку можна використати тезаурус, у цьому разі до слів запиту додаються близькі їм відповідники із тезаурусу [3]. Це дозволить знаходити релевантні текстові фрагменти, які взагалі не місять зазначених ключових слів. Разом із цим ступіть відповідності знайденого матеріалу інформаційному запиту зменшується за рахунок наявності в Інтернеті різного виду замаскованої інформації, зміст якої не відповідає головному заголовку. Метою такої діяльності є підвищення рейтингу власників сайтів за рахунок збільшення звернень від користувачів у відповідь на сенсаційну інформацію. У результаті подібного пошуку оператор отримує великі обсяги зайвої інформації, яка не відповідає заданій тематиці. Можливим шляхом розв'язання даної проблеми є створення бази даних синонімів, яка дозволить частково автоматизувати процес формування інформаційного запиту. Крім того, АСПІ повинна самонавчатися. Зробити це можливо завдяки аналізу результату пошуку (перевірки відповідності знайденого матеріалу заданій фразі). Оператору має бути надана можливість обрати варіант пошукової фрази, яка формується з отриманого завдання-речення.

Проведений аналіз наявних АСПІ показав, що для ефективного пошуку тематичної інформації в мережі Інтернет запит повинен формуватися з урахування [6]:

можливості використання логічних зв'язків;

обмеження у відстані між ключовими словами запиту в документі;

пошуку за визначеними полями документа;

можливості призначення вагових коефіцієнтів словам у запиті;

проведення сортування за датою, розміром документа тощо.

Ранжування знайдених документів повинне відбуватися на основі міри близькості "запит – документ". Один з ефективних методів ранжування полягає в перевірці частоти появи ключового слова запиту у відповідних полях HTML-документа: <Title>, <H1…H6>, <ADDRESS>, <B>, <STRONG>. Чим ближче розташовано слово до початку документа, тим більший його ваговий коефіцієнт. Для встановлення вагових коефіцієнтів окремим пошуковим словам оцінюється їх наявність у заголовку та підзаголовку документа.

Майбутня АСПІ повинна враховувати різноманітні служби Інтернету, що привертають увагу аналітиків спеціальних служб, такі як: пошукові системи, тематичні каталоги ресурсів, сайти новин, RSS-повідомлення та інформаційні агентства, які транслюють новини онлайн. Для досягнення цієї мети першим кроком у розробці АСПІ є створення бази даних пошукових ресурсів, які доступні в Інтернеті, з урахуванням їхніх особливостей щодо надання інформації за визначеною тематикою. Для цього пропонуємо використовувати такі рубрики в базі даних (табл. 1) [7].

Таблиця 1

Рубрики бази даних пошукових ресурсів

№ з/п	*Назва пошукового ресурсу*	*Адреса ресурсу*	*Рейтинг за результатами власника*	*Кількість запитів*	*Кількість знайдених посилань за запитом*	*Ступінь відповідності запиту (ступінь релевантності)*	*Частота оновлення інформації за визначеною тематикою*
1	2	3	4	5	6	7	8

Уточнення пошуку можливе в разі використання тематичної класифікації ресурсів – векторів простору словника (термінів індексації) системи [8]. При цьому завдання полягатимуть у тому, щоб найкращим чином сформувати правила та обрати такі ознаки, на основі яких буде прийматися рішення щодо віднесення ресурсу до певної рубрики. Тематичні рубрики повинні бути контекстонезалежними. Використання створеної бази даних перед безпосереднім пошуком дозволить оператору сформувати метапошуковий запит з урахуванням апріорних даних, що будуються на оцінюванні результатів пошуку з конкретного ресурсу.

Окремо розглянуто завдання щодо вилучення інформації з онлайн відеотрансляції новин. У цьому разі для своєчасного реагування оператору необхідно постійно відслідковувати визначені завчасно передачі. Одним із можливих варіантів виконання цього завдання в умовах обмеженої кількості особового складу є здійснення запису трансляцій та формування плану їх перегляду. При цьому можна скористатися інтернет-порталами, що надають інформацію про розклад телепередач. Це дозволить автоматично обирати контент, який відповідає тематичному пошуку.

Подальша робота АСПІ пов'язана з обробкою відібраного матеріалу. Для цього необхідним етапом у ході створення майбутньої системи є організація автоматичного реферування знайденої інформації. Порядок оброблення документа складається з певних етапів (рис. 2).

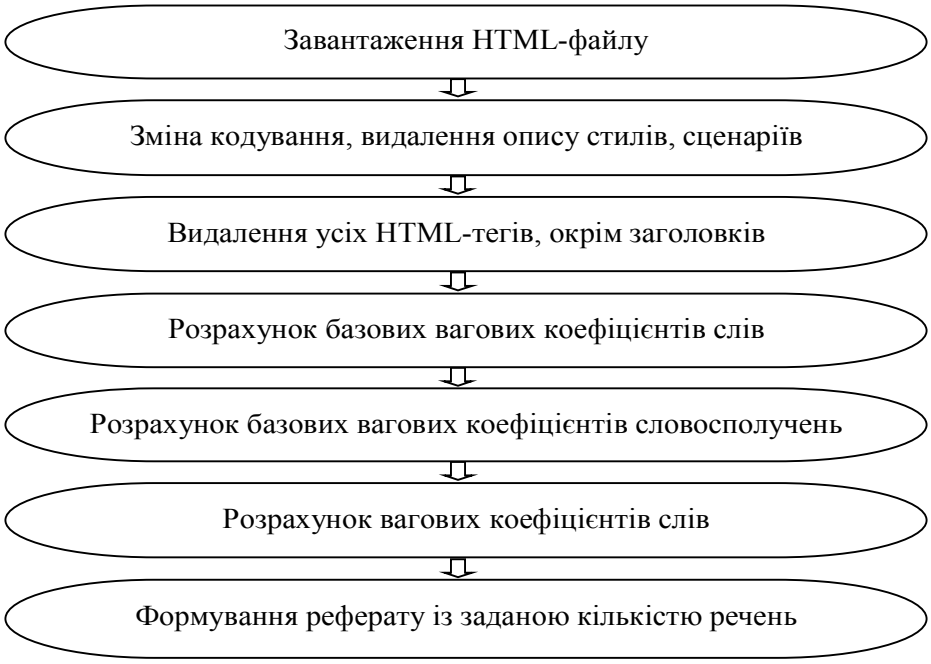

Рис. 2. Етапи автоматичного реферування документів

Для візуалізації знайденої інформації з метою її подальшого аналізу доцільно використати технологію побудови семантичних мереж [8]. Порівняння семантичних мереж різних текстів дозволяє встановити ступінь їх змістової близькості, що може використовуватися для автоматичної класифікації документів за заданими рубриками, їх пошуку за подібністю заданого тексту, а також розбиття інформаційного масиву на класи документів близького змісту. Структурна схема АСПІ за визначеною тематикою в мережі Інтернет матиме такий вигляд (рис. 3).

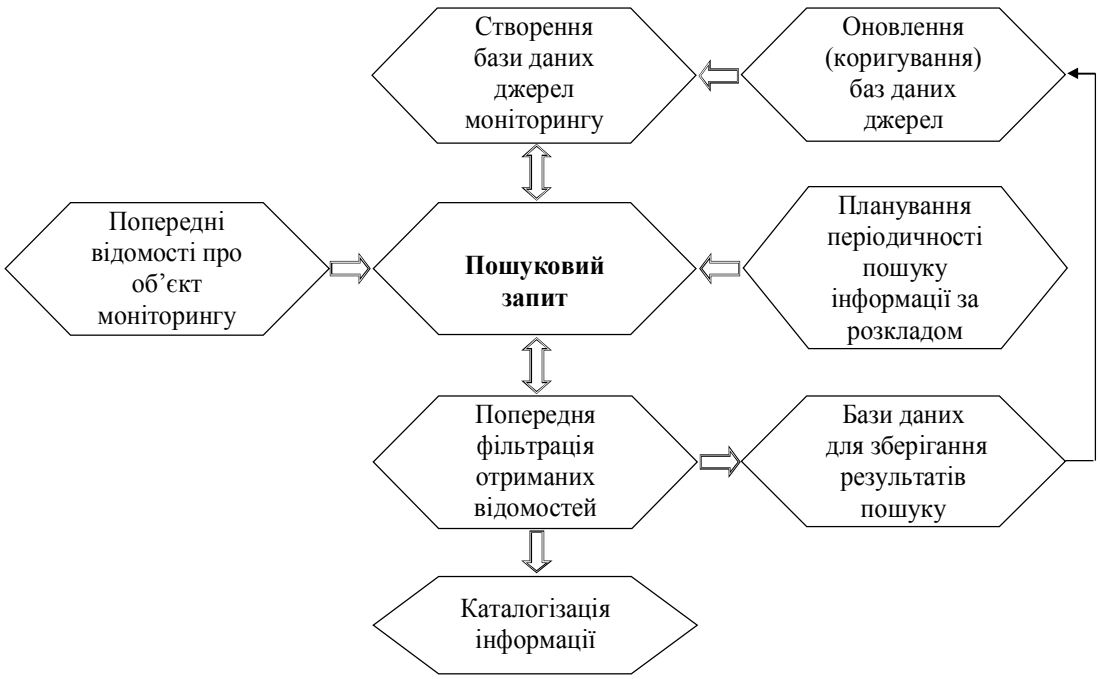

Рис. 3. Структурна схема АСПІ

Висновки. Реалізація процесу автоматизованого пошуку тематичної інформації в мережі Інтернет за запропонованою структурною схемою під час інформаційно-аналітичної діяльності в спеціальних підрозділах можлива шляхом:

забезпечення оператора-аналітика засобами швидкого та ефективного пошуку різнорідної інформації за об'єктами моніторингу;

надання засобів швидкого виявлення неявних зв'язків між об'єктами моніторингу та пов'язаними з ними фактами й подіями;

проведення фіксації та візуалізації результатів аналітичних досліджень шляхом генерації дайджестів статей, фактів, формалізованих досьє, семантичних мереж й інших аналітичних звітів.

Перспективи подальших досліджень у даному напрямі полягають у розробленні архітектури відповідного спеціалізованого програмного забезпечення.

СПИСОК БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ

1. Nych L. Y., Kaminskyj R. M., & Shakhovska N. B. Effectiveness evaluation of search in information systems with consolidated information // Radio Electronics, Computer Science, Control. 2016. https://doi.org/10.15588/1607-3274-2016-2-13

2. Поташова А. В. Проблеми пошуку інформації в глобальній мережі Інтернет // Науковий огляд. 2018. № 5 (48). С. 130–139.

3. Рогушина Ю. В. Методика розробки термінології інформаційних ресурсів як базису формування онтологій та тезаурусів для семантичного пошуку // Інженерія програмного забезпечення. 2014. № 1. С. 41–51. URL: http://nbuv.gov.ua/UJRN/Ipz_2014_1_7 (дата звернення: 15.02.2023).

4. Крайовський В. Я., Литвин В. В., Шаховська Н. Б. Основні підходи до розроблення програмного комплексу автоматичного реферування текстових документів // Зб. наук. праць Ін-ту проблем моделювання в енергетиці ім. Г. Є. Пухова НАН України. Київ : ІПМЕ ім. Г. Є. Пухова НАН України, 2009. Вип. 51. С. 178–186.

5. Грищук Р. В., Даник Ю. Г. Основи кібернетичної безпеки : Монографія. Житомир : ЖНАЕУ, 2016. 549 с.

6. Системний аналіз інформаційних процесів : навч. посіб. / В. М. Варенко, І. В. Братусь, В. С. Дорошенко та ін. Київ : Ун-т «Україна», 2013. 203 с.

7. Міхєєв Ю. І., Чернявський Г. П., Манько О. В., Токар А. М. Контент-моніторинг у системі виявлення інформаційних небезпек // Зб. тез доп. Міжнар. наук.-техн. конф. "Перспективи розвитку озброєння та військової техніки сухопутних військ" (14–16 травня 2014). Львів : АСВ, 2014. С. 39.

8. Сухий О. Л., Міленін В. М., Тарадайнік В. М. Алгоритми пошуку в інформаційних системах : методич. рекомендації. Київ, 2015. С. 70.

Стаття надійшла до редакції 30.11.2023.

REFERENCES

1. Nych, L. Y., Kaminskyj, R. M., & Shakhovska, N. B. (2016). Effectiveness Evaluation of Search in Information Systems with Consolidated Information. *Radio Electronics, Computer Science, Control.* https://doi.org/10.15588/1607-3274-2016-2-13

2. Potashova, A. V. (2018). Problemy poshuku informatsii v hlobalnii merezhi Internet [Problems of Information Search in the Global Internet]. *Naukovyi ohliad [Scientific Review], 5* (48), 130–139 [in Ukrainian].
3. Rohushyna, Yu. V. (2014). Metodyka rozrobky terminolohii informatsiinykh resursiv yak bazysu formuvannia ontolohii ta tezaurusiv dlia semantychnoho poshuku [Methodology for Developing the Terminology of Information Resources as a Basis for the Formation of Ontologies and Thesauri for Semantic Search]. *Inzheneriia prohramnoho zabezpechennia [Software Engineering], 1,* 41–51. Retrived from http://nbuv.gov.ua/UJRN/Ipz_2014_1_7 [in Ukrainian].
4. Kraiovskyi, V. Ya., Lytvyn, V. V., & Shakhovska, N. B. (2009). Osnovni pidkhody do rozroblennia prohramnoho kompleksu avtomatychnoho referuvannia tekstovykh dokumentiv [Basic Approaches to the Development of a Software Complex for Automatic Abstracting of Text Documents]. *Zb. nauk. prats In-tu problem modeliuvannia v enerhetytsi im. H. Ye. Pukhova NAS Ukrainy [Collection of Scientific Papers of the Pukhov Institute of Modeling Problems in Energy of the National Academy of Sciences of Ukraine], Iss. 51,* 178–186. Kyiv [in Ukrainian].
5. Hryshchuk, R. V., & Danyk, Yu. H. (2016). *Osnovy kibernetychnoi bezpeky : Monohrafiia [Fundamentals of Cyber Security: Monograph].* Zhytomyr [in Ukrainian].
6. Varenko, V. M., Bratus, I. V., & Doroshenko, V. S. et al. (2013). *Systemnyi analiz informatsiinykh protsesiv : navch. posib. [Systematic Analysis of Information Processes: Textbook].* Kyiv [in Ukrainian].
7. Mikhieiev, Yu. I., Cherniavskyi, H. P., Manko, O. V., & Tokar, A. M. (2014). Kontent-monitorynh u systemi vyiavlennia informatsiinykh nebezpek [Content Monitoring in the System of Detection of Information Threats]. In *Zb. tez dop. Mizhnar. nauk.-tekhn. konf. "Perspektyvy rozvytku ozbroiennia ta viiskovoi tekhniky sukhoputnykh viisk" [Collection of Abstracts of the International Scientific and Technical Conference "Prospects for the Development of Weapons and Military Equipment of the Ground Forces"*. Lviv, May14–16, 2014. (pp. 39–40). Lviv [in Ukrainian].
8. Sukhyi, O. L., Milenin, V. M., & Taradainik, V. M. (2015). *Alhorytmy poshuku v informatsiinykh systemakh : metodych. Rekomendatsii [Search Algorithms in Information Systems: Methodological Recommendations].* Kyiv [in Ukrainian].

Y. I. Mikhieiev, V. V. Loboda, T. M. Voitko, R. I. Hladych

AUTOMATED SYSTEM OF SEARCHING FOR INFORMATION ON INTERNET

In this article, author examines a process of searching for information on a particular topic on the Internet. Attention is focused on features related to functioning an information retrieval system, and the operation of its typical model is studied. Such a system should provide a flexible response to an operator's request by means an interactive search and maintenance a query archive, including a thesaurus, spelling and punctuation checkers for a language query. Results of analysis carried out on existing automated information retrieval systems indicate that a query should be formed taking into account: using logical connections, restrictions on the distance between words, the possibility of setting weighting coefficients for keywords, sorting by date and document size.

An approach to creating and a structural scheme of an automated system of searching for thematic information on Internet based on unique nature of operational and analytical activities of special units is proposed. Such a system should take into account particularities in operation of various Internet services used by special services analysts: thematic resource catalogs, news sites, RSS feeds and news agencies broadcasting news online. It is envisaged that development of special software based within the proposed functional structure of automatic system for searching thematic information on Internet will increase efficiency of information and analytical activities in special units due to availability of means for: searching for heterogeneous information on specific objects; identifying links between the monitored objects and related facts and events; visualization of analytical research results.

Keywords: *automated system; Internet; information query; information retrieval system; structural diagram.*