УДК 6081.518:004.451.6

**Н.О. Ткаченко (аспірант), В.Я. Воропаєва (канд. техн. наук, доц.)**
ДВНЗ «Донецький національний технічний університет», м. Донецьк
науково-технічна бібліотека ДонНТУ,
кафедра автоматики і телекомунікацій
E-mail: nsot@library.dgtu.donetsk.ua, voropayeva@meta.ua

# АЛГОРИТМ РОБОТИ ІНФОРМАЦІЙНО-ПОШУКОВОЇ СИСТЕМИ ЗІ ЗВОРОТНИМ ЗВ'ЯЗКОМ

*У статті висвітлено актуальну проблему оптимізації інформаційного пошуку серед великої кількості електронних ресурсів. Проведено аналіз та класифікацію пошукових систем зі зворотним зв'язком. Розроблено критерії оцінки релевантності інформації залежно від потреб користувача. Запропоновано структуру та алгоритм роботи нової адаптивної інформаційно-пошукової системи зі зворотним зв'язком.*
***Ключові слова:*** *пошукова система, інформаційні потреби користувача, системи зі зворотним зв'язком, релевантність.*

**Вступ**

Ускладнення інформаційних вимог наукової комунікації призвело до ситуації, коли друковані фонди традиційних бібліотек вже не в змозі забезпечувати потреби користувачів в інформації. Бібліотеки наукових установ та університетів активно створюють електронні колекції. Крім того, всесвітня мережа Інтернет щогодини поповнюється різноманітними електронними документами, що розміщуються на різних платформах. Це призвело до виникнення проблеми релевантного пошуку інформації окремим користувачем, бо традиційні пошукові системи використовують автоматичні алгоритми пошуку. Ці алгоритми надають однакову інформацію усім користувачам, враховуючи лише персональні оцінки веб-сайтів [1]. Проблема пошуку релевантної інформації стосується не лише глобальних пошукових систем, локальні пошукові системи окремих країн, та навіть навчальні заклади, що мають електронні колекції, потерпають від цієї ж самої проблеми. Розташування електронних ресурсів на різних платформах знижує якість пошуку інформації, бо призводить до підвищення часових витрат з боку користувача: користувач повинен повторювати свій запит в різних пошукових інтерфейсах. Але навіть такий спосіб пошуку інформації не гарантує отримання релевантної інформації, бо ігнорує інформаційні інтереси окремого користувача. Отже, розробка єдиної оптимальної системи пошуку інформаційних ресурсів, адаптованих до потреб окремого користувача, все ще залишається актуальною задачею.

**Постановка проблеми**

Донецький національний технічний університет також має власну електронну колекцію, що оновлюється кожного місяця. Окрім електронних підручників та методичних вказівок вона нараховує велику кількість наукових ресурсів як власної генерації, розташованих в основному в інституціональному репозитарії відкритого доступу (ea.donntu.edu.ua), так і отриманих із зовнішніх джерел (тестові та передплачені бази даних або електронні журнали), а також ресурсів, що розміщуються на сайтах окремих кафедр.

Важливою частиною електронної колекції ДонНТУ є інтелектуальна система дистанційного навчання, що дозволяє підвищити якість освіти завдяки адаптації системи під конкретного учня [2].

Велику популярність останнім часом завойовують ресурси відкритого доступу, що виступили потужним конкурентом традиційним платним журналам. Ідея відкритого доступу

швидко отримала підтримку вчених потужних навчальних та наукових закладів світу. На фоні цього у ДонНТУ на базі бібліотеки було створено електронний архів (репозитарій) відкритого доступу, що зберігає наукові матеріали вчених університету. Це покращує ступінь наукової комунікації вчених університету, робить його частиною світового наукового простору та дозволяє підвищити позиції навчального закладу у різноманітних рейтингах, зокрема WEBOMETRIX [3].

З іншого боку, наявність великої кількості розрізнених електронних ресурсів у бібліотеці університету дозволяє стверджувати про необхідність розробки єдиної системи пошуку релевантних документів.

Велику популярність в останні роки завойовують пошукові системи зі зворотним зв'язком. Залежно від виду зворотнього зв'язку такі системи умовно розділяють на (рис.1): системи з явним зворотним зв'язком (systems with explicit relevance feedback) та системи з неявним зворотнім зв'язком (systems with implicit relevance feedback) [4].

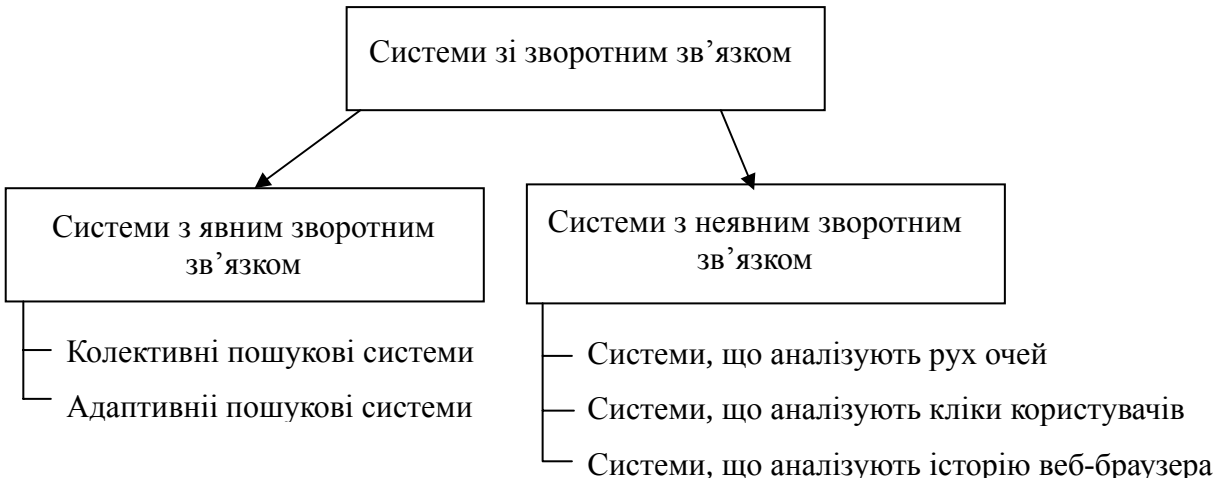


Рисунок 1 - Класифікація пошукових систем зі зворотним зв'язком

Усі пошукові системи з неявним зворотнім зв'язком аналізують неявну поведінку користувачів у всесвітній мережі. Зворотний зв'язок з користувачем відбувається неявно, тобто без його прямої участі. Оскільки усі ці системи враховують поведінку користувача, то їх класифікація відбувається умовно, в залежності від того, що вони обробляють.

Системи, що аналізують рух очей (eye tracking).

Аналіз активності користувачів при відборі релевантної інформації, що надала пошукова система, довів, що вибір багатьох користувачів прямо пропорційно залежить від популярності сайту. Тобто частіше за все користувачі зупиняють свій вибір на перших двох десятках результатів, навіть якщо вони не містять релевантних документів. Авторами [5] було проведено дослідження щодо залежності переваг користувачів від довжини анотації сайту. Дослідження складалося з двох частин: навігаційне та інформаційне. Усі запити оброблялися на окремому сервері, рух очей відстежувався за допомогою додаткового обладнання. Авторами отримано висновок, що переваги користувачів при пошуку інформації залежить від довжини анотації сайту: сайтам з короткою анотацією користувачі віддають більшу перевагу.

Системи, що аналізують кліки користувачів.

T.Joachims [6] використовує SVM (Support Vector Machine) алгоритм для аналізу кліків користувачів. Згідно автору, ці клікі надаються у вигляді трійки *(q, r, c)*, що складається з *q* запитів, *r* документів, наданих користувачеві, та *c* посилань, на які користувач натиснув. Для алгоритму використовується функція Кендала: для запиту *q* та колекції документів $D=\{d_1,\ldots,d_m\}$ оптимальна пошукова система повинна повертати документи в порядку їх

релевантності до інтересів користувача. Для двох різних рейтингів документів $r_a$ та $r_b$ функція Кендала залежить від двох параметрів: кількості пар, що співпадають *(P)* та кількості різних пар *(Q)*. Пара документів $d_i \neq d_j \in$ P, якщо обидва рейтинги документів $r_a$ та $r_b$ розташували їх в однакову порядку, в іншому випадку пара відноситься до *Q*. Для цього випадку функція Кендала обчислюється за наступною формулою:

$$\tau(r_a, r_b) = \frac{P-Q}{P+Q} = 1 - \frac{2Q}{2m}. \quad (1)$$

Такий підхід раціональний лише для бінарної оптимізації. Для фіксованої, але невідомої послідовності запитів $P_r(q, r^*)$ та рейтингів ранжування колекції D з $m$ документами, отримання оптимальної функції ранжування можливе при максимізації функції Кендала:

$$\tau_p(f) = \int \tau(r_{f(q)}, r^*) dP_r(q, r^*). \quad (2)$$

Не дивлячись на позитивні результати, що були отримані авторами описаних систем, вони все ж таки мають один великий недолік — це статичність, тобто ці системи не встигають за змінами в інформаційних потребах користувача. Тому останнім часом все більшу популярність завойовують адаптивні системи, що здатні пристосовуватися до змін у інтересах користувача.

У роботі [7] Kazunari Sugiyama та ін. розроблено новий метод пошуку релевантної інформації, який дає змогу системі адаптуватися до змін в інтересах користувачів завдяки аналізу історії веб-браузера. Головна відмінність цієї системи від вже існуючих у тому, що аналізується не лише довгострокова історія веб-браузера, а й поточна історія (тобто історія веб-браузера за поточний день). Це дозволяє враховувати актуальні зміни в інформаційних потребах користувача та оновлювати його інформаційний профіль. Формула для створення або оновлення такого профілю виглядає наступним чином:

$$P = \alpha P^{per} + \beta P^{today} = \alpha P^{per} + \beta x P^{br} + \beta y P^{cur}, \quad (3),$$

де $P^{per}$ - історія відвіданих сторінок за декілька днів, $P^{br}$ - історія відвіданих сторінок за поточний день та $P^{cur}$ - поточна сесія пошуку інформації.

Завдяки аналізу неявної поведінки користувача в мережі системи з неявним зворотним зв'язком хоч і дозволяють опосередковано підвищити результати пошуку релевантних документів, все ж таки мають окремий недолік — поведінка користувача у мережі не завжди вказує на його власні інформаційні інтереси.

На фоні цієї проблеми почали розвиватися пошукові системи з явним зворотним зв'язком, що враховують інтереси користувачів шляхом їх залучення до процесу пошуку інформації у всесвітній мережі. До таких систем відносяться, зокрема, рекомендуючі системи, в яких переваги групи користувачів відіграють вирішальну роль у наданні релевантної інформації. Сайти отримують рекомендації за допомогою персональних оцінок користувачів. Для організації таких систем використовуються два різних підходи - спільна фільтрація та системи рекомендації контенту:

- спільна фільтрація (Collaborative Filtering**).** Для пошуку релевантних документів системи на основі спільної фільтрації використовують дані попередніх оцінок інших користувачів [8];
- системи рекомендації контенту порівнюють зміст знайденої сторінки з контентом, що цікавить окремого користувача.

Рекомендуючі системи надають релевантні результати при наданні користувачами добровільних оцінок запропонованим ресурсам. Це створює додаткові навантаження на окремого користувача, тому зазвичай такі рекомендації не можуть виступити гарантом релевантності через небажання користувачів витрачати зайвий час на оцінку ресурсу.

У роботі [9] авторами запропоновано адаптивну пошукову систему зі зворотним

зв'язком. Для пошуку інформації система використовуватиме онтологічний профіль користувача, що будується методом аналізу історії веб-браузеру користувача, а додатковий інтерфейс дозволить користувачу приймати участь у відборі інформації. Окрема онтологія являє собою орієнтований граф, у якому кожному поняттю та зв'язкам між поняттями присвоюються індивідуальні ваги. А кінцева вага окремої онтології розраховується за формулою:

$$W = \sum \omega * \varphi, \quad (4),$$

де $\omega$ - вага окремого поняття онтології, $\varphi$ - вага зв'язків між цими поняттями.

Вага кожного поняття в документі розраховується за відомою tf * idf формулою:

$$\omega_{ij} = tf_{ij} * idf_i, \quad (5),$$

де $tf = \frac{m_i}{\sum m_p}$ - частота слова, $idf = \log \frac{S}{P}$ - зворотна частота документу.

Наявність великої кількості електронних ресурсів в Донецькому національному технічному університеті викликає необхідність створення єдиної інформаційно-пошукової системи, яка оброблятиме запити користувачів та надаватиме релевантну інформацію. Система, що розробляється, матиме змогу адаптуватися до змін в інтересах користувачів та використовуватиме явним зворотний зв'язок для уточнення користувачем своїх запитів. Для виконання зазначених функцій система складається з чотирьох модулів (рис.2).

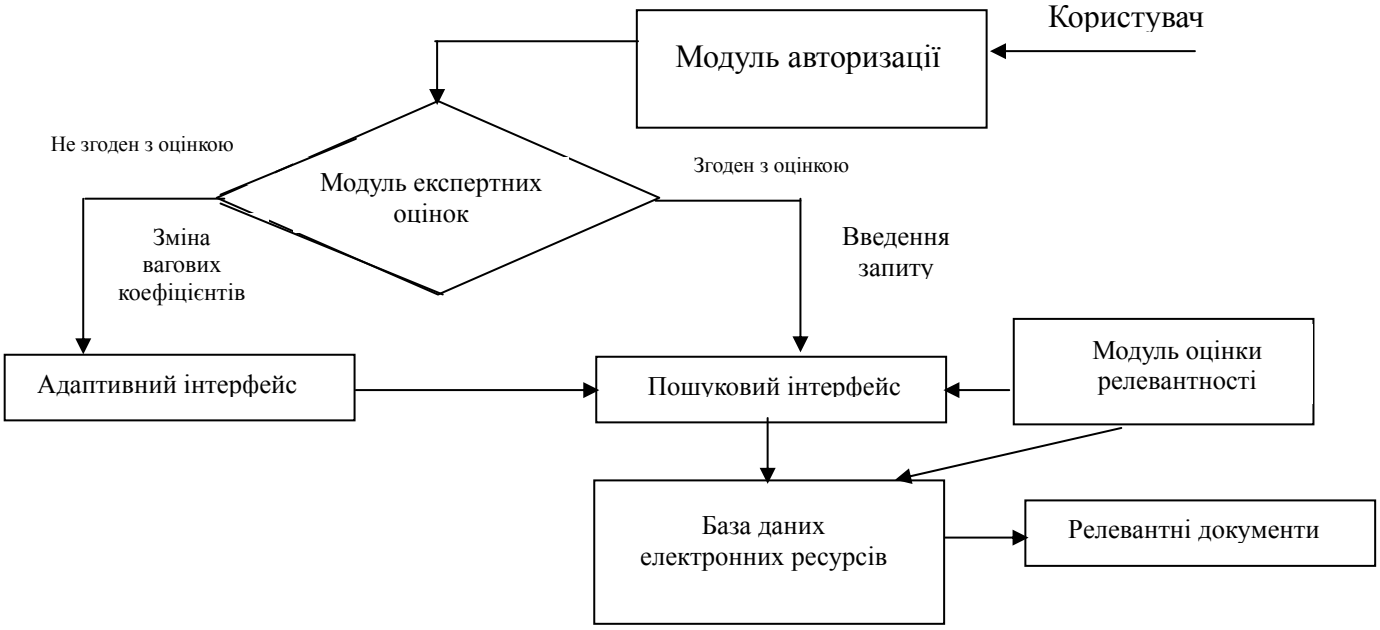


Рисунок 2 - Структура інформаційно-пошукової системи

Модуль авторизації містить усі дані про користувачів інформаційно-пошукової системи, які розділяються на п'ять категорій: адміністратор, бібліотекар, викладач, студент та молодий вчений. Тобто власне споживачі інформації умовно поділені на 3 групи, залежно від їх переважних інтересів. За умовчанням прийнято, що студенти переважно шукають у бібліотеці навчальні матеріали, молоді вчені – наукові ресурси, а викладач може цікавитися як навчально-методичними, так і науковими джерелами.

Адміністратор виконує функції налаштування, управління, усунення несправностей в системі та надання відповідних прав іншим категоріям користувачів. Співробітники бібліотеки оновляють базу даних електронних ресурсів: додають до бази нові електронні ресурси та видаляють вже не актуальні. Усі інші категорії користувачів отримують права на пошук інформації в системі та, при необхідності, її збереження і роздруківки.

Модуль електронних ресурсів об'єднує розподілені електронні ресурси в єдину базу даних, серед якої здійснюється пошук інформації. До складу цієї бази даних входять: власні електронні ресурси бібліотеки, електронний архів ДонНТУ, матеріали інших репозитаріїв та журналів відкритого доступу згідно тематики університету, передплачені та тестові ресурси, а також матеріали системи дистанційної освіти.

Модуль експертних оцінок містить у своєму складі оцінки ресурсів для різних груп користувачів, що визначаються відповідними експертами та завантажуються у систему. Для кожної категорії визначається експертна оцінка ресурсів, яка складається з декількох критеріїв з певними ваговими множниками. Але користувач має змогу змінювати ці множники відповідно до своїх інформаційних потреб, що дозволяє системі адаптуватися під конкретного користувача.

Модуль оцінки релевантності, ґрунтуючись на функції релевантності, здійснює пошук релевантної інформації серед бази даних електронних ресурсів.

Алгоритм роботи системи.

1. На першому етапі читач авторизується у системі за допомогою номеру свого читацького квитка. Система аналізує дані читача та розподіляє його до однієї з категорій користувачів.
2. Для введення свого запиту користувач використовує спеціальний інтерфейс, що підтримує зворотний зв'язок з читачем.
3. На цьому етапі система звертається до модулю експертних оцінок та обирає ту, що відповідає обраній категорії користувача. Вона пропонує користувачеві визначені вагові множники для кожного з критеріїв. Якщо користувач згоден, система переходить до іншого етапу. Якщо ні – користувач особисто змінює значення вагових множників відповідно до своїх інформаційних інтересів та повертає їх системі.
4. Система звертається до бази даних електронних ресурсів та за допомогою обраної функції релевантності повертає користувачеві релевантні документи.

**Висновки**

У статті надана класифікація інформаційно-пошукових систем зі зворотним зв'язком. Приведені основні переваги та недоліки систем з явним та неявним зворотним зв'язком. Запропоновано принцип роботи адаптивної пошукової системи зі зворотним зв'язком. Авторами розроблено структуру та алгоритм роботи нової інформаційно-пошукової системи з явним зворотним зв'язком, яка відрізняється від існуючих тим, що дозволяє користувачеві особисто приймати участь у пошуковому процесі для підвищення релевантності пошуку.

**Список використаної літератури**

1. Ткаченко, Н.О. Розробка критеріїв для оцінки інформаційних наукових ресурсів в Інтернет / Н.О. Ткаченко// Наукові праці Донецького національного технічного університету. Серія: Обчислювальна техніка та автоматизація. - 2013. - Вип. 2 (25). - С. 136-143.
2. Воропаєва, В.Я. Математичне моделювання процесів дистанційного навчання / В.Я Воропаєва, Д.В. Криворучко // Научно-технический журнал «Автоматика. Автоматизация. Электротехнические комплексы и системы». – 2004. - № 2(14). - С. 11-15.
3. Воропаєва, В.Я Вплив електронного архіву ВНЗ на наукометричні показники / В.Я. Воропаєва, Н.О. Ткаченко // Бібліотеки та інформаційні ресурси у сучасному світі науки, освіти та культури : Матеріали наук. - практ. конф., м. Севастополь, 7-10 жовтня 2013 р. Севастополь: Купол , 2013 . С. 13-15.
4. Kelly Diane Implicit Feedback for Inferring User Preference: A Bibliography / Diane Kelly, Jaime Teevan// SIGIR. – 2003. - №37. - С. 18-28.
5. Cutrell, E. What are you looking for? An eye-tracking study of information usage in Web search / E. Cutrell, G. Zhiwei // Proceedings of the SIGCHI conference on Human factors in

computing systems. – 2007. – С. 407-416.
6. Joachims, T. Optimizing Search Engines using Clickthrough Data. / T. Joachims, // Proceedings of the ACM Conference on Knowledge Discovery and Datamining. – 2002. С. 184–203.
7. Kazunari Sugiyama Adaptive Web Search Based on User Profile Constructed without Any Effort from Users / Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa // Proceedings of the 13th international conference on World Wide Web. - 2004. - С. 675 – 684.
8. Goldberg, D. Using Collaborative Filtering to Weave an Information Tapestry / D.Goldberg, D. Nichols, B.M. Oki, D.B. Terry // Communications of the ACM. - 1992. - 35(12). – С.61–70.
9. Ткаченко, Н.О. Принципи функціонування інформаційно-пошукової системи зі зворотним звязком / Н.О. Ткаченко, В.Я. Воропаєва, М.М. Дученко // Збірник наукових праць ДонІЗТ. - 2013.- № 36.- С. 83-90.

**References**

1. Tkachenko, N.A. (2013), "Development of the Criteria to Evaluate Scientific Information Resources in Internet", *Naukovyi praci Donetsk National Technical University.Ser.Obchusluvalna tehnika ta avtomatuzaciya* , vol. 2, no. 25, pp. 137-143.
2. Voropayeva, V.Y. and Krivoruchko, D.V. (2004), "Mathematical modeling of Distance Learning", Nauchno-tehnicheskiy jornal "Avtomatika.Avtomatizaciya.Electrotehnicheskie kompleksu I systemu", vol. 2, no. 14, pp. 11-15.
3. Voropayeva, V.Y. and Tkachenko, N.A. (2013), "Effect of electronic university archive on Scientometrics indicators", *Proceedings of the scientific conference Libraries and Information Resources in the Modern World of Science, Education and Culture*, Sevastopol, Ukraine, 7-10 October 2013, pp. 13-15.
4. Kelly, Diane and Teevan, Jaime (2003), "Implicit Feedback for Inferring User Preference: A Bibliography", ACM SIGIR Forum, vol. 37, no. 2, pp. 18-28.
5. Cutrell, E. and Zhiwei, G. (2007), "What are you looking for? An eye-tracking study of information usage in Web search", ACM, *Proceedings of the SIGCHI conference on Human factors in computing systems*, California, USA, 28 April - 3 May, 2007, pp. 407-416.
6. Joachims, T. (2002), "Optimizing Search Engines using Clickthrough Data", ACM, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, USA, 23-26 April 2002, pp. 184–203.
7. Kazunari, Sugiyama, Kenji, Hatano and Masatoshi, Yoshikawa (2004), "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users", *Proceedings of the 13th international conference on World Wide Web*, New York, USA, 17-22 May 2004, pp. 675-684.
8. Goldberg, D., Nichols, D., Oki, B.M. and Terry, D.B. (1992), "Using Collaborative Filtering to Weave an Information Tapestry", *Communications of the ACM*, vol. 35, December, pp. 61–70.
9. Tkachenko, N.A., Voropayeva, V.Y. and Duchenko, M.M. (2013), "*Principles of adaptive information retrieval system with feedback", Zbirnik Prac' DonIZT*, No. 26, pp. 83-90.

Надійшла до редакції:
28.04.2014 р.

Рецензент:
докт. техн. наук, проф. Скобцов Ю.О.

***Н.А. Ткаченко, В.Я. Воропаева***
***ГВУЗ «Донецкий национальный технический университет»***

***Алгоритм работы информационно-поисковой системы с обратной связью.*** *В статье рассмотрена актуальная проблема информационного поиска среди огромного количества электронных ресурсов. Был проведен анализ и предложена классификация поисковых систем с обратной связью. Разработаны критерии оценки релевантности информации в*

*зависимости от потребностей пользователя. Предложены структура и алгоритм работы новой адаптивной информационно-поисковой системы с обратной связью.*
***Ключевые слова:*** *поисковая система, информационные потребности пользователя, системы с обратной связью, релевантность.*

***N.A. Tkachenko, V.Y. Voropayeva***
***Donetsk National Technical University***
***Information retrieval system with feedback operation algorithm.*** *This article describes the problem of information search. Detailed analysis of existing retrieval systems with feedback was given. The authors proposed a classification of such systems: systems with explicit feedback and systems with implicit feedback. Systems with explicit feedback attract users to the process of finding relevant information, i.e. user indicates its information needs individually. These systems include: collective and adaptive systems. In contrast, a system with implicit feedback analyze only user behavior on the worldwide web (eye movement, the history and the history of clicks user's web browser), which facilitate the burden on the user. Adaptive systems take up a special place among systems with feedback, they are able to adapt to changes in the information the user's interests. Analysis of the electronic collection in Donetsk National Technical University was given in the article and the necessity of creating a search engine among these electronic resources. Authors developed an adaptive information retrieval system with feedback, which includes four modules: an authorization module, expert evaluation module, the database of electronic resources and assessment relevance module. Authorization module contains information about all users of the system that allows to divide them into several categories: administrator, librarian, student, a young scholar and teacher. Electronic resources module combines all electronic resources of the University. Expert evaluation module contains expert estimation of resources for each category of user. Assessment relevance module contains relevance function working algorithm, by which relevant information among database of electronic resources is selected. The algorithm of this system is presented in the article.*
***Keywords:*** *retrieval system, information needs of the user, system with feedback, relevance.*

**Ткаченко Наталія Олександрівна**, Україна, закінчила Донецький національний технічний університет, бібліотекар НТБ ДонНТУ, аспірант кафедри автоматики та телекомунікацій ДВНЗ «Донецький національний технічний університет» (вул. Артема, 58, м. Донецк, 83001, Україна). Основний напрямок наукової діяльності - оптимізація сучасних пошукових систем.

**Воропаєва Вікторія Яківна**, Україна, закінчила Донецький національний технічний університет, канд. техн. наук, доцент, професор кафедри автоматики та телекомунікацій ДВНЗ «Донецький національний технічний університет» (вул. Артема, 58, м. Донецк, 83001, Україна). Основний напрямок наукової діяльності – сучасна теорія телетрафіку, оптимізація телекомунікаційних та інформаційно-комунікаційних систем та мереж.