УДК 004.5

О. Г. Додонов, Д. В. Ланде, В.Г. Путятін
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна,

В. В. Жигало
ІЦ «Електронні Вісті»
вул. М. Кривоноса, 2-А, 03037 Київ, Україна

Архітектура системи моніторингу, адаптивного агрегування та узагальнення інформації

Представлена функціональна схема узагальненої системи моніторингу, адаптивної агрегації та узагальнення потоків інформації з глобальних комп'ютерних мереж для забезпечення інформаційно-аналітичної діяльності. Наданий опис основних етапів обробки інформації та інформаційних потоків, наведено стислий опис макету – експериментальної системи. Практичне значення роботи полягає в обгрунтуванні підходів і засобів створення інформаційно-аналітичного середовища для проведення науково-аналітичних досліджень.

Ключові слова: *функціональна схема, моніторинг інформації, метапошукова система, адаптивна агрегація інформації*

Аналітична діяльність передбачає роботу з інформацією, її глибоке осмислення, прийняття рішень з аналізу тієї чи іншої ситуації, отримання додаткової інформації, аналіз всієї наявної інформації, що відноситься до проблеми, тематичну обробку інформації, підготовку та візуалізацію аналітичних звітів, їх верифікацію, отримання управлінських рішень на базі нових знань [1].

Таким чином, нагальним для аналітика є своєчасне отримання об'єктивної документальної інформації, у тому числі за допомогою засобів моніторингу комп'ютерних мереж, сучасних пошукових і метапошукових систем для подальшого її використання у своїх дослідженнях.

Розвиток інформаційних мережевих технологій привів до значного зростання об'ємів документальної інформації в мережевому середовищі. Не дивлячись на те, що велика кількість аналітичних матеріалів публікується на «закритих» інформаційних ресурсах (тих, які вимагають оплати, реєстраційних даних, корпоративної приналежності і

т.п.), велика частина з них публікується у веб-середовищі. Разом з тим, зростання об'єму інформаційного середовища супроводжується багатократним дублюванням інформації, слабкою її структуризацією, зростанням рівня інформаційного шуму [2,3].

Тому особливо актуальною є розробка теоретичних і технологічних принципів побудови адаптивних інформаційних сховищ, автоматизованих систем обробки і узагальнення інформації з документальних сховищ надвеликого об'єму, які повинні стати основою для створення інтелектуального середовища рішення аналітичних міждисциплінарних проблем.

Таким чином, для підтримки інформаційно-аналітичної діяльності необхідно реалізувати теоретично обґрунтовані технології, що охоплюють всі ланцюжки роботи з інформацією, включаючи моніторинг, агрегування та узагальнення потоків інформації.

Одним із підходів до агрегування інформаційних потоків можна вважати створення інформаційно-пошукових систем – програмного забезпечення, призначеного для пошуку і відображення документів у базах даних, які зможуть акумулювати документи з цих потоків [4]. Представляється дуже важливим, щоб агрегація інформації та формування інформаційного сховища були адаптивними, тобто орієнтованими на інформаційні потреби користувачів [5]. Якщо враховувати динаміку і об'єми доступної інформації в Інтернеті (на сьогодні доступний понад трильйон документів), то стає очевидним, що забезпечення ефективного доступу в режимі пошуку до інформації у відриві від інформаційних потреб, є практично нерозв'язним завданням.

Якщо ранжирувати кількість джерел, які необхідні в аналітичній діяльності, ймовірно, можна у черговий раз отримати підтвердження наукометричної закономірності Бредфорда [6], яка, у свою чергу витікає із закону Ципфа. Закономірність Бредфорда у початковому вигляді відносилася до традиційних «паперових» періодичних видань. Досліджуючи різні типи джерел інформації, Бредфорд розподілив їх за трьома множинами, рівними за кількістю релевантних документів: R_1, R_2, R_3. При цьому R_1 – це найбільш рейтингові джерела, які безпосередньо відносяться до певної тематики; R_2 – множина джерел, що кореспондуються з комп'ютерною тематикою; R_3 – джерела, які лише частково торкаються даної теми. При цьому кількість корисної інформації в усіх трьох множинах є сталою.

Якщо прийняти позначення, що $|A|$ - це кількість елементів множини A, то пропорція Бредфорда записується у такий спосіб:

$$|R_1| : |R_2| : |R_3| = C.$$

Для множин документальних джерел, що отримуються за наведеними вище алгоритмами, відповідно, справедливим є вираз:

$$|S_1| : |S_3| = |S_3| : |S_2| = C,$$

де S_1 – це множина заздалегідь відомих аналітику джерел, отриманих з веб-простору,

S_2 – множина джерел, що отримані шляхом пошуку в глобальних мережевих пошукових системах;

S_3 – джерела, що отримуються шляхом застосування спеціалізованих метапошукових систем;

C – деяка константа, що відповідає інформаційним потребам користувачів.

Основна ідея адаптивної агрегації інформації полягає у зборі і збереженні в інформаційному сховищі тільки тієї інформації, яка відповідає інформаційним потребам користувачів (існуючих або потенційних). Для цього передбачається, що по мірі розвитку системи (і придбання популярності) в її інформаційне сховище потраплятимуть актуальні документи з Інтернету, відповідні поточним запитам користувачів. Природно, із зростанням кількості користувачів, об'єми інформаційного сховища (репозитарія) будуть також зростати, що у конкретний момент призведе до перегляду його вмісту за деякими критеріями, наприклад, за змістом, використовуючи методи Text Mining, або за часом відповідно до формули Бартона-Кеблера [7]:

$$m(t) = 1 - ae^{-T} - b\,\mathrm{e}^{-2T},$$

де $m(t)$ – частина корисної інформації через час T, перший від'ємник відповідає стабільним ресурсам (наприклад, монографіям, фундаментальним звітам), а другий – динамічним (наприклад, тези доповідей, новини).

Існує багато спільного між існуючими на даний час системами агрегування інформації, але слід зауважити, що багато із розглянутих можливостей є унікальними з погляду на ідеологію і технологію кожної окремої системи.

В результаті проведених досліджень і узагальнень пропонується модель системи моніторингу, адаптивного агрегування та узагальнення інформації, функціональну схему якої наведено на рис. 1.

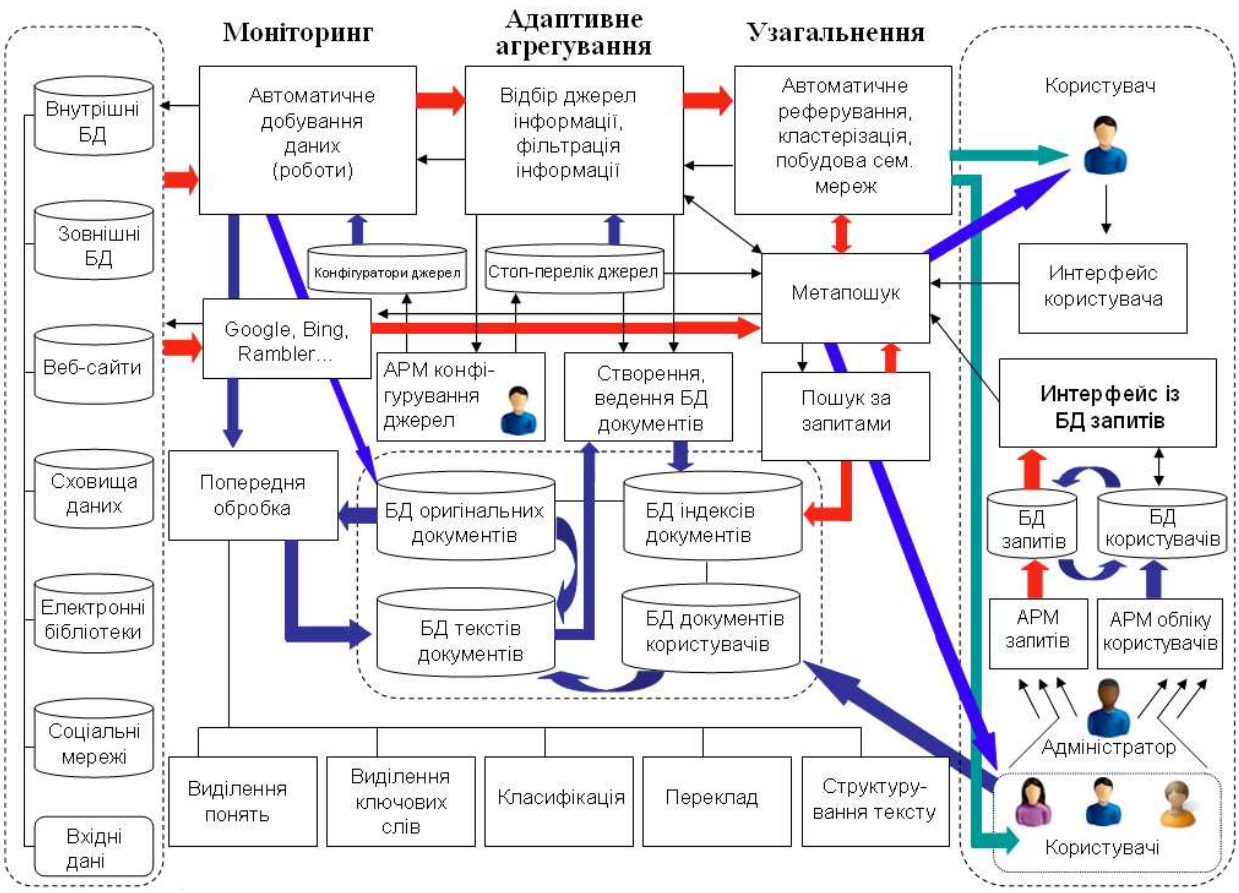

Рис. 1 – Функціональна схема системи моніторингу, адаптивного агрегування та узагальнення інформації

У відповідності до цієї схеми, окремий користувач звертається до системи через спеціальний інтерфейс. Передбачається, що користувач може бути зареєстрованим, тоді він має більші права, зокрема, може управляти власним інформаційним кешем, чого не може робити незареєстрований користувач системи (відвідувач). Крім того, постійні статичні запити окремих користувачів можуть вводитися для обробки у підсистемі вибіркового розповсюдження інформації (засобами електронної пошти) через адміністратора. Адміністратор вводить запити користувачів за допомогою спеціалізованого автоматизованого робочого місця (АРМ). Для обліку користувачів він також застосовує окреме програмне забезпечення, за допомогою якого ведеться база даних користувачів. Вміст бази даних запитів користувачів через спеціальний прикладний інтерфейс передаються метапошуковій системі, до якої також передаються запити користувачів, що працюють з системою в режимі онлайн.

Всі зареєстровані користувачі мають можливість розміщувати свої власні документи у базу даних, до якої реалізовано доступ у пошуковому режимі (також через

метапошукову систему). Тобто ядром взаємодії користувача з системою є метапошукова система [8, 9], яка по суті агрегує зовнішні і внутрішні інформаційні ресурси і надає користувачу доступ до них через «єдине вікно».

Метапошукова система забезпечує взаємодію з зовнішніми мережевими пошуковими системами, такими, як, наприклад, Google, Bing, Rambler, за допомогою яких реалізується пошук у таких інформаційних джерелах, як веб-сайти, деякі соціальні мережі, зовнішні бази даних, сховища даних, архіви, електронні бібліотеки тощо.

Одним з головних принципів побудови моделі адаптивного документального сховища є принцип фільтрації результатів. Він полягає у тому, що по-перше, відбувається фільтрація неінформативних сайтів або сайтів з недоступними першоджерелами (так званий «чорний список», або «стоп-перелік»). Крім того, адаптивна метапошукова система розбирає отримані результати на окремі документи і перевіряє їх доступність. Наприклад, якщо у шляху до документу присутнє доменне ім'я, присутнє в «стоп-переліку», то документ відкидається і не використовується у подальшій обробці. Це лише один з критеріїв фільтрації. Ті документи, які пройшли етап фільтрації, перетворяться для виведення результатів користувачу. Також здійснюється пошук у внутрішній базі цих файлів (у інформаційному кеші на проксі-сервері, що містить знайдені раніше документи). Якщо такі файли були знайдені, то виведення документу доповнюється інформацією про можливу доступність цього файлу за знайденим посиланням. Якщо цей файл відсутній за вказаною адресою в Інтернеті, то виводиться повідомлення, що цей файл може бути відсутнім. Якщо ж інформація про цей файл присутня в інформаційному кеші і він імовірно існує, то вивід доповнюється інформацією, такою, як розмір файлу, а також створюється HTML-версія цього файлу. Після підрахунку кількості знайдених документів підготовлені результати виводяться користувачу через стандартний веб-інтерфейс.

Таким чином, користувачу надаються лише ті документи, які пройшли спеціальну фільтрацію. Фільтри створюються, з одного боку, інформаційним адміністратором, який формує «стоп-перелік» джерел інформації, доступ до яких потребує передплати, реєстрації, містить лише метадані щодо документів і т.д., а, з другого боку, програмними застосуваннями, що не дозволяють видавати посилання на неіснуючі документи, або переправляють посилання до кешу системи, де цільові документи вже розміщені у результаті опрацювання попередніх запитів.

Другий принцип побудови адаптивного документального сховища полягає на налаштуванні на вже знайдені користувачами документи. Тобто реалізується модуль кешування, основне завдання якого – збір посилань на документи, які отримані в процесі роботи з користувачем метапошукової системи, щоб надалі зберегти в інформаційному

сховищі (кеші системи) файли, а також пов’язану з ним інформацію, таку, як доступність файлу по цьому посиланню і розмір файлу.

Система періодично оновлює інформацію про ті файли, які були збережені в базі даних. Якщо файл не був раніше доступний, але доступний у той момент, коли виробляється вторинне сканування, інформація в базі даних оновлюється; якщо ж він стає недоступним, то в базу даних записується інформація про недоступність цього файлу, щоб у подальшому запропонувати користувачу отримати цей файл з кешу. Знайдені і відмічені користувачами як корисні документи підлягають попередній обробці, зокрема, вони переводяться у текстовий формат, структуруються для відображення (за бажанням користувачів) у цьому форматі. Крім того, шляхом екстрагування з документів виділяються окремі поняття (персони, компанії, топоніми тощо), за лінгво-статистичними критеріями виділяються ключові слова, здійснюється переклад ключових слів іншими мовами, класифікація документів за тематиками. Вибрані документи завантажуються до баз даних оригінальних документів, куди також (у необхідному обсязі) завантажуються у режимі автоматичного моніторингу документи з вибраних адміністратором сховищ даних, електронних бібліотек, архівів. Кожний оригінальний документ супроводжується своїм текстовим образом у базі даних текстових документів. Відібрані поняття і ключові слова завантажуються у базу даних індексів документів, до яких і звертається внутрішня пошукова система.

В результаті реалізації функціональної схеми, що розглядається, користувач за своїми запитами може отримувати як переліки доступних релевантних документів, так і узагальнені звіти, дайджести, переліки сюжетних ланцюжків, інтерактивні семантичні мережі, діаграми трендів понять або подій тощо.

Основним критерієм ранжирування інформації в сучасній метапошуковій системі має бути рейтинг пошукових систем. Так, наприклад, у пошукової системи Google рейтинг вищий, ніж у системи Bing (у Google больше охоплення ресурсів, більш релевантні результати). Якщо посилання на один і той же PDF-документ було отримано з різних пошукових систем, то вибирається те з них, яке містить найбільш повний опис.

Таким чином, запропонована узагальнена метапошукова система реалізує такі необхідні етапи обробки мережевої інформації, як моніторинг, адаптивне агрегування (за інформаційними потребами користувачів) та узагальнення інформації, що створює передумови для реалізації пошукового середовища нового типу для підтримки інформаційно-аналітичної діяльності.

Як прототип узагальненої системи моніторингу, адаптивного агрегування та узагальненні інформаційних потоків авторами запропонована модельне рішення – система

Doc's Bundle, яка дозволяє шукати документи у форматі PDF як в Інтернеті, так і в спеціально накопиченому кеші документів (усередині системи) в процесі роботи. Формат PDF як основний для модельного рішення було обрано тому, що нині в інтернет-просторі знаходиться велика кількість документальних ресурсів, представлених у цьому форматі. Популярність цього формату обумовлена тим, що він є компактним і зручним для зберігання інформації, представленої з самого початку у вигляді простого тексту, векторних і растрових зображень, сторінок веб-сайтів, форм і мультимедійних файлів. В той же час, при пошуку необхідної документації у форматі PDF за допомогою традиційних мережевих інформаційно-пошукових систем користувач постійно стикається з проблемами, пов'язаними з поганою доступністю цільової інформації (умовами платного доступу, відсутністю необхідних файлів по вказаних адресах, або невірними гіперпосиланнями). Хоча більшість пошукових систем, таких як Google, Yandex, Rambler, Yahoo і ін. виводять в список результатів інформацію про знайдені документи, в той же час вони часто дають посилання на неіснуючі PDF-файли, або посилання на веб-сайти, де документи знаходяться в закритому доступі.

Під час виконання пошуку користувач системи Doc's Bundle потрапляє на сторінку (рис. 2), де представлені результати пошуку з декількох пошукових систем, які були вибрані ним у списку. Також користувачу пропонується список ключових слів для швидкої навігації в системі. Документ в списку представлений такими даними:

- назва документа з посиланням на джерело;
- анотація документа;
- посилання на скачування документа з кешу системи, якщо даний документ потрапив у кеш, розмір файлу;
- допоміжна інформація про те, в якій пошуковій системі документ було знайдено.

Для того, щоб користувач отримав максимальні можливості системи, він повинен увійти в систему або зареєструватися. Для входу в систему користувач має ввести логін і пароль, після чого йому надаються розширені можливості:

- завантаження власних файлів у кеш для зберігання і пошуку в ньому;
- відбирати файли в «Обране», а також пошук у «Обраному»;
- створення власних рубрик.

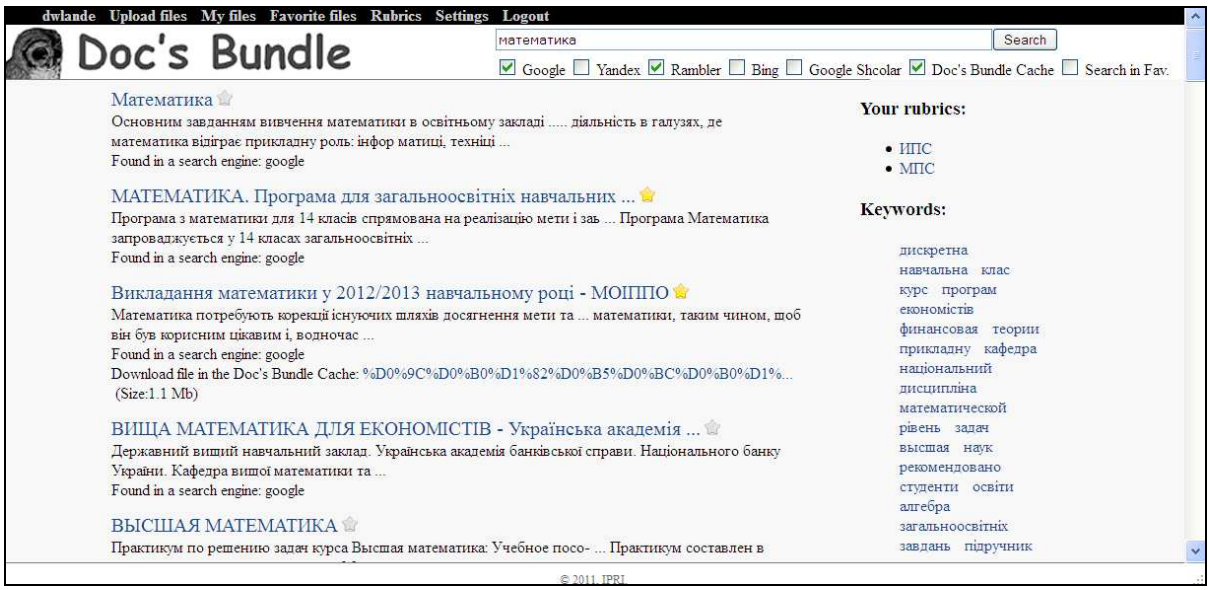

Рис. 2 – Сторінка пошуку PDF-документів

На сторінці пошуку документів користувачу доступно (рис. 2):

– меню користувача, яке дозволяє використовувати розширені функції системи;

– форма пошуку з додатковим параметром, за допомогою якого можна проводити пошук в «Обраному»;

– список документів;

– блок користувальницьких рубрик;

– блок ключових слів для навігації користувача по системі.

Документ у списку представлений такими даними:

– назва документа з посиланням на джерело;

– спеціальний інструмент «Зірочка» за допомогою якого користувач може додати/видалити документ в «Обране»;

– анотація документа;

– посилання на скачування документа з кешу системи, якщо даний документ потрапив у кеш, розмір файлу;

– допоміжна інформація щодо того, в якій інформаційно-пошуковій системі даний документ було знайдено.

Зареєстрований користувач може закачувати в систему свої документи в форматі PDF, а також проводити пошук в закачаних файлах.

Користувальницькі файли приймають участь в пошуку і для інших користувачів, у випадку якщо був обраний пошук у кеші системи.

Файли, додані користувачем у систему автоматично додаються до списку вибраних документів. У разі, якщо користувач зняв з файлу відмітку «Обраний» - файл через деякий час може бути видалений.

Вибрані файли – додатковий функціонал користувача за допомогою якого він може відкласти корисний для нього файл і надалі не проводити додатковий пошук.

Щоб додати або видалити документ в «Обране» користувач повинен натиснути на спеціальний елемент інтерфейсу «Зірочка». Якщо зірочка жовта - документ доданий до вибраного, сіра – документ не в обраному.

Користувач може переглянути весь список обраних файлів на сторінці Favorite files або пошукати за запитом у «Обраному» зазначивши в пошуковій формі параметр Search in Fav.

Всі файли, додані у «Обране», не видаляються з кешу системи, і існують до тих пір, поки існує хоч один користувач який додав файл в «Обране». Крім того, зареєстрований користувач може додавати, змінювати і видаляти користувацькі рубрики (рис. 3). На сторінці пошуку користувачу представлений список його рубрик у вигляді окремого блоку.

Рис. 3 – Сторінка додавання користувацьких рубрик

Розглянута модель вже нині знайшла своїх користувачів і дозволила сформулювати складніші завдання, які мають бути вирішені при побудові корпоративних інформаційно-аналітичних систем [9, 10]. Передбачається, що результати проведених досліджень складуть теоретичну базу для розробки автоматизованих систем моніторингу, адаптивної агрегації і узагальнення інформаційних документальних потоків, побудови і ведення інформаційних ресурсів надвеликих об'ємів і різноманітною тематичній спрямованості, дозволять поєднати в єдиному технологічному ланцюжку моніторинг, інформаційний пошук, агрегація інформації із змістовним аналізом даних, їх узагальненням, що

підвищить якість обробки інформації з глобальних мереж, і, відповідно, ефективність інформаційно-аналітичної підтримки науково-аналітичної діяльності вітчизняних учених і фахівців.

Література

1. *Додонов О.Г., Путятін В.Г., Валетчик В.О.* Інформаційно-аналітична підтримка прийняття управлінських рішень // Реєстрація, зберігання і обробка даних. – 2005. – **7**.– № 2. – С. 77-93.

2. *Додонов А.Г., Ландэ Д.В., Жигало В.В.* Сетевые информационные потоки как содержательная составляющая информационно-аналитических систем // Реєстрація, зберігання і обробка даних, 2010. – 12. – № 1. – С. 39-48.

3. *Додонов О.Г., Ланде Д.В., Путятін В.Г.* Інформаційні потоки в глобальних комп'ютерних мережах. – К: Наукова думка, 2009, – 295 с.

4. *Lande D., Braichevski S., Busch D.* Informationsfluesse im Internet // IWP - Information Wissenschaft & Praxis, 2007. – **5**. – № 59 – P. 277-284.

5. *Додонов А.Г., Ландэ Д.В.* Методы и средства мониторинга, адаптивного агрегирования и обобщения информационнных потоков // Информационные технологии и безопасность. Проблемы научного и правового обеспечения кибербезопасности в современном мире. Материалы международной научной конференции ИТБ-2011. – К.: ИПРИ НАН Украины, 2011. – С. 6-9.

6. *Bradford S. C.* Sources of information on specific subjects // Journal of Information Science, 10:4, 1985 (October). – P. 173–180.

7. *Bruton R., Kebler R.* The half-life of some scien-tific and technical literature // Am. Document, 1960. – 11. – № 1. – P. 18-22.

8. *Meng W., Yu C, Liu K.L.* Building Efficient and Effective Metasearch Engines // ACM Comput. Surv. 34, 1 (Mar. 2002). – P. 48-89.

9. *Ландэ Д.В., Снарский А.А., Жигало В.В.* Метапоиск доступных научно-технических документов в Интернет // Труды 12^{й} Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010. – С 321-325.

10. *Ландэ Д.В., Снарский А.А.* Возможности системы мультипоиска доступных научно-технических документов в Интернет на примере тематики неразрушающего контроля и технической диагностики // Матеріали 15-ої міжнародної науково-технічної конференції "електромагнітні та акустичні методи неруйнівного контролю матеріалів та виробів", 15-20 лютого 2010 р., Славське Львівської обл. – С. 105-107.