УДК 681.3

І.Ю. Гришанова

АНАЛІТИЧНИЙ ОГЛЯД МЕТОДІВ І ЗАСОБІВ ІНФОРМАЦІЙНОГО ПОШУКУ В SEMANTIC WEB

В статті надаються та аналізуються методи і засоби інформаційного пошуку в середовищі Semantic Web. Надаються базові поняття інформаційного пошуку, задачі, моделі та класифікація систем інформаційного пошуку за різними ознаками. Наводяться приклади існуючих сучасних пошукових систем, а також надається перелік ознак 3-х поколінь пошукових систем. Запропонована модель інформаційного пошуку в новому середовищі Semantic Web та Web речей розширює класифікацію пошукових систем та модель пошуку з урахуванням можливості пошуку нових об`єктів, доступних по інтернету, та використання знань, що подані в Semantic Web.

Ключові слова: інформаційний пошук, семантичний пошук, пошукові системи, Semantic Web.

Вступ

Філософське і історичне визначення інформаційного пошуку. Важливість персоніфікації у процесі пошуку

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].

Інформаційний пошук є процес знаходження матеріалу (зазвичай документів) неструктурованої природи (частіше текстів), які задовольняють інформаційній потребі, у великих колекціях (зазвичай, які зберігаються на комп`ютерах).

Класичне поняття інформаційного пошуку (IR – *information retrieval*, ІП) базується на задоволенні потреби користувачів у пошуку інформації, тобто інформаційної потреби (information need). Класичне визначення інформаційного пошуку базується на підставі того факту, що користувач спонукається інформаційною потребою.

В найбільш загальному сенсі під ***інформаційною потребою*** розуміється необхідність в інформації, яка потребує задоволення і зазвичай виражена в інформаційному запиті. Наприклад, планування поїздки формує інформаційну потребу вивчити розклад руху поїздів та іншого транспорту. Такий процес може бути виконаний різним чином – за допомогою телефону, безпосередньо в касах, в агенстві з продажу квитків, або за допомогою пошуку та сайту в Інтернеті. Однак незалежно від форм задоволення інформаційної потреби, сама по собі вона залишається невідмінною.

Необхідно зазначити, що коли необхідний маршрут обрано та білети вже придбані, ця інформація втрачає свою цінність для користувача, при цьому вона залишається цінною для інших потенційних споживачів. Така властивість повної втрати цінності інформації (її споживацької вартості) для певного споживача в певний момент, є важливою особливістю інформаційної потреби, що суттєво відрізняє її від інших видів потреб людини. Одна й та сама інформація знов може стати предметом споживання в випадках, якщо вона буде надана іншому споживачеві, або якщо перед тим самим споживачем знов стане така сама задача, або якщо запас знань споживача зросте, що дозволить йому побачити в цій інформації нові аспекти.

Таким чином, інформаційні потреби мають суто індивідуальний (персональний) характер. Вони залежать не тільки від особливостей задач, що вирішуються, але й від психологічних, освітніх та інших особистих відмінностей особи, що приймає рішення.

Зазвичай виділяють два основних типи інформаційних потреб:

- поточні, які зумовлені притаманною людині допитливістю і які виражаються в його прагненні бути в курсі усьо-

го, що відбувається в світі;

- конкретні (спеціальні), які виражаються в прагненні отримати інформацію, необхідну для вирішення конкретної задачі – дослідницької, професійної, управлінської тощо [2].

Основна мета задачі інформаційного пошуку – допомогти користувачу знайти інформацію, яка йому необхідна. Процес інформаційного пошуку в загальному вигляді включає в себе послідовність операцій, які направлені на збір, обробку і надання необхідної інформації зацікавленим особам. Процес інформаційного пошуку складається з наступних етапів:

- визначення (уточнення) інформаційної потреби і формулювання інформаційного запиту;
- визначення сукупності можливих інформаційних джерел;
- вилучення інформації з виявлених інформаційних джерел;
- ознайомлення з отриманою інформацією і оцінювання результатів пошуку.

Базовими поняттями оцінювання ефективності пошуку є **релевантність** та **пертинентність.**

Вирішальною умовою ефективного задоволення інформаційної потреби є чітке усвідомлення і чітке вираження того, яка інформація насправді потрібна споживачеві для вирішення поставленого перед ним завдання. Без цього важко розраховувати на отримання релевантного та пертинентного результату.

З моменту виникнення у людини інформаційної потреби, він починає оцінювати всю інформацію, що надходить до нього, під кутом зору цієї потреби, розділяючи цю інформацію на релевантну і нерелевантну. Іншими словами, інформаційна потреба виникає у людини при постановці перед нею якогось завдання. Людина обмірковує цю задачу, в результаті чого у нього в мозку складається образ задачі, або її модель. Цей образ і служить еталоном, з яким порівнюється вся подальша інформація, що надходить. Якщо інформація має відношення до еталону, вона вважається доречною. Все, що не має відношення до еталону – вважається **нерелевантною** інформацією.

Під впливом міркувань над сутністю поставленої задачі та вмістом релевантної інформації, що накопичується, уява людини про цю задачу може уточнюватися та змінюватися. Психологи називають такий процес зростанням стану поінформованості про завдання.

Коли людиною накопичено необхідну кількість інформації і виконано деякий міркувальний процес, вона знаходить рішення задачі. Після цього вся інформація, що пов’язана з рішенням задачі, переміщується в зону архівного зберігання. Таким чином, інформаційна потреба може бути охарактеризована як усвідомлена потреба в інформації, яка необхідна для вирішення поставленої задачі за розробленим планом.

Можливо припустити, що процес вирішення будь-якої наукової задачі починається з прийняття будь-яких передумов і припущень, які в подальшому піддаються коригуванню і зміні. Під образом чи моделлю завдання слід розуміти гіпотезу, яка є важливим засобом організації наукового пошуку.

Вчення про психологічні установки дозволяє пояснити поняття пертинентності, яке є одним з ключових понять теорії інформаційного пошуку. Під **пертинентністю** розуміється відповідність знайдених документів або відомостей справжній інформаційній потребі вченого або спеціаліста, яку він нерідко сам може ясно не усвідомлювати.

З запропонованої інтерпретації сутності інформаційної потреби та механізму її задоволення випливає, що віднесення інформації, що надходить до людини, до категорії релевантної чи нерелевантної, повністю визначається тим, який образ поставленої задачі склався у даної людини. Сам цей образ залежить, принаймні, від трьох наступних факторів:

- інформації, яка вже накопичена людиною в її пам’яті;
- обраного шляху рішення задачі;
- темпів і проміжних результатів рішення.

Ще раз необхідно зазначити, що образ завдання під впливом інформації, що надходить, та проміжних результатів рішення цієї задачі, уточнюється або навіть змінюється. У зв'язку з цим змінюються і ознаки, за якими розпізнається і відбирається релевантна інформація. Тому для адекватного інформаційного обслуговування фахівців необхідно, щоб процес пошуку був не тільки індивідуальним, але й включав у себе постійний зворотній зв'язок для своєчасного урахування змін у його інформаційній потребі.

Базові поняття інформаційного пошуку

Основним засобом передачі інформації у часі і просторі є документ. ***Документ*** визначається як засіб закріплення будь-яким чином на спеціальному матеріалі будь-якої (деякої) інформації про факти, події, явища об'єктивної дійсності і розумової діяльності людини [3]. Документи мають різну форму подання. В автоматизованих інформаційно-пошукових системах це текстова інформація на природній мові. В повсякденному житті – це може бути друкована стаття, книга тощо. В Інтернет це може бути рисунок, відео-ролик або сайт.

З точки зору теорії інформації ***документ*** – це змістовно закінчена одиниця інформації, яка представлена на якій-небудь природній мові, що ідентифікується унікальним чином.

Поняття інформаційного пошуку вперше запровадив в інформатиці американський математик Келвин Муерс в 1947 році. ІП називається деяка послідовність операцій, яка виконується з метою знаходження документів, які містять певну інформацію (з подальшою видачею цих документів або їх копій), або з метою видачі фактичних даних, які надають відповіді на задані питання.

Спонукальним приводом інформаційного пошуку, як було зазначено вище, є інформаційна потреба, яка виражена у формі інформаційного запиту. Об'єктами інформаційного пошуку можуть бути документи, відомості про їх наявність та/або місцезнаходження, фактографічна інформація.

Інформаційний запит представляє собою інформаційну потребу, яка сформульована на природній мові. Результат «перекладу» інформаційного запиту на інформаційно-пошукову мову (ІПМ) називають ***пошуковим образом запиту*** (ПОЗ). Синтаксис і семантика ІПМ визначається структурою і наповненням документів та загальними задачами системи.

Інформаційний пошук розрізняють наступним чином:

- в залежності від мети – адресний пошук (формально-механічний) та семантичний (тематичний);
- в залежності від об'єкта пошуку – документний та фактографічний;
- в залежності від ступеня використання технічних засобів – ручний або автоматизований;
- в залежності від функціональної ролі – домінуючі/другорядні, центральні /периферичні, сталі/ситуаційні потреби.

Усі види інформаційного пошуку перетинаються, тому що цілі та об'єкти часто взаємопов'язані. Наприклад, документний і фактографічний види пошуку можуть бути як адресними, так і семантичними.

Інформаційний пошук здійснюється за допомогою інформаційно-пошукових систем. ***Інформаційно-пошукова система*** (ІПС) – це комплекс пов'язаних між собою окремих частин, який призначений для виявлення в будь-якій множині елементів інформації, які відповідають заданому інформаційному запиту. Масив елементів інформації, в якому виконується інформаційний пошук, називається *пошуковим масивом*.

Інформаційно-пошукові системи розділяються на *документальні* та *фактографічні*. Документальні ІПС у відповідь на запит видають оригінали, копії або адреси місцезнаходження документів, що містять потрібну інформацію. Підклас документальних ІПС, які видають лише бібліографічні описи документів, що знайдені, іноді називаються бібліографічними ІПС.

На відміну від документальних ІПС фактографічні пошукові системи призначені для видачі безпосередньо необхідної інформації (наприклад, температури кипіння якоїсь рідини, температури води в морі біля конкретного населеного пункту; структурних або молекулярних формул хімічних сполук, що мають певні властивості тощо).

Принципової відмінності між документальними і фактографічними ІПС немає. Головною ознакою, що поєднує документальні і фактографічні ІПС до одного загального класу є те, що на запити вони можуть видавати таку й тільки таку інформацію, яка була раніше в них введена.

Кожна документальна ІПС (як ручна, так і автоматизована), містить наступні частини:

- ІПМ;
- правила перекладу текстів документу і запитів з природної мови на ІПМ;
- формальні правила (алгоритми) пошуку;
- технічні засоби, які реалізують алгоритми пошуку;
- масив (множина) документів (або їх адрес), які записані на якихось носіях інформації (в сучасних пошукових системах Інтернету – база індексу).

Інформаційний пошук здійснюється за певними правилами, які визначають стратегію пошуку, тобто способи досягнення оптимального результату. Стратегія інформаційного пошуку залежить від типу пошукової задачі, критеріїв видачі і характеру діалогу між споживачами інформації і ІПС.

В загальному вигляді процедура інформаційного пошуку складається з чотирьох етапів:

- уточнення інформаційної потреби і формулювання запиту;
- визначення сукупності інформаційних масивів;
- вилучення інформації з інформаційних масивів;
- ознайомлення користувача з отриманою інформацією і оцінювання результатів пошуку.

Найбільш загальний вигляд алгоритму пошуку, що проводиться незалежно від форми носіїв і ступеня автоматизації, показаний на рис. 1.

Рис. 1. Загальний вигляд алгоритму пошуку

Постановка пошукової проблеми. На цьому етапі користувач формулює точне визначення і фіксує те, що буде шукати і в якій області знань (предметній області – ПрО). Таким чином множина пошуку звужується визначеними межами.

Створення тезаурусу проблеми. На цьому етапі користувач створює (складає) перелік слів, які найбільш повно відображають ПрО або проблему, що була визначена. Як рекомендують спеціалісти з бібліографічного пошуку, цей перелік повинен мати приблизно 10–15 слів.

В залежності від поставленого завдання тезаурус може бути складений на декількох мовах, для пошуку серед вітчизняних та зарубіжних джерел інформації. Робота над тезаурусом ведеться весь час, і в процесі виявлення нових термінів вони

тут же додаються до тезаурусу. Найбільш прийнятною є структура тезаурусу у вигляді семантичних зрізів. У цьому випадку для кожного основного терміну окремо будується таблиця для супутних та шумових слів. Шумових слів у джерелі бути не повинно. Тобто користувач отримує пакет таблиць, які можна окремо розширювати і модифікувати в ході пошуку.

Відбір джерел інформації для пошуку. Джерела інформації (масив) обираються виходячи з характеру проблеми (тобто де найбільш доступні та повно надані джерела) та можливостей користувача (доступ до Інтернету, бібліотеки тощо).

Виконання пошуку засобами, які притаманні джерелу інформації. На цьому етапі користувач з тезаурусу складає пошукові запити і реалізує їх методами пошуку, які специфічні для даного ресурсу. В бібліотеці – це пошук в каталогах, якщо інформацією володіють люди або організації – пошук та звернення до них, у мережі Інтернет – використовуються пошукові машини та каталоги, телеконференції та списки розсилки, сайти та інше. Як формат, так і семантика запитів варіюється в залежності від предметної області та використовуваного інформаційного ресурсу.

Як рекомендують спеціалісти з бібліографічного пошуку, запити необхідно складати таким чином, щоб область пошуку була максимально конкретизована та звужена. Необхідно віддавати перевагу декільком вузьким запитам ніж одному, але розширеному. В загальному випадку для кожного основного поняття з тезаурусу готується окремий пакет запитів. Після чого проводиться пробне виконання запитів – для уточнення та доповнення тезаурусу, в тому числі для відсікання шумової інформації.

Оцінювання отриманих результатів пошуку. В результаті пошуку користувач отримує результативну множину документів, які надалі необхідно проаналізувати і вирішити наскільки повно вони покривають поставлену пошукову проблему.

Перелік ресурсів, отриманих у результаті запиту, рекомендується обробляти в два етапи. На першому етапі відсікаються вочевидь нерелевантні джерела і знову ж таки проводиться семантичний аналіз з метою уточнення тезаурусу та модифікації подальших запитів. На другому етапі обробки користувач послідовно вивчає кожен з знайдених ресурсів для безпосереднього аналізу інформації, що знаходиться в них. У процесі аналізу отриманої інформації, її треба:

- оцінити (за ступенем вірогідності, важливості, таємності, пов'язаності між собою, можливості використання);
- інтерпретувати (в світлі інших даних і глибинної інтуїції), виявивши її місце в загальній мозаїці фактів;
- визначити, чи потрібна (і яка) додаткова інформація;
- ефективно використати (врахувати у своїх планах, передати кому слід, притримати до потрібного моменту).

Прийняття рішення про продовження (закінчення) пошуку. Якщо, оцінюючи результати пошуку, користувач прийшов до висновку, що необхідна інформація знайдена вся, тоді пошук можна припиняти – подальші пошуки будуть зайвою тратою дорогоцінного часу. У зворотній ситуації (неповні відомості) користувачеві доведеться приймати рішення про те, на якому з етапів була допущена помилка, і спробувати виправити її, після чого повторити процес пошуку з цього місця заново. В цьому випадку можливі три варіанти: неправильно складений тезаурус проблеми, неправильно обране інформаційне джерело або користувач скористався недоцільними методами пошуку (наприклад, виконував пошук суто наукової інформації – статті за допомогою загально використовуваного пошукового Інтернет-сервісу). Такі ітерації необхідно повторювати, поки не буде досягнуто позитивного результату. При цьому існує стовідсотково методологічна проблема – при ефективному пошуку завжди стоять два суперечливих завдання: збільшення охоплення з метою отримання максимальної кількості значимої інформації та зменшення охоплення з метою мінімального обсягу шумової інформації. І най-

складніше, як завжди, знайти золоту середину [4].

Найбільш ефективним методом пошуку документів, які містять наукову інформацію є вивчення (прочитання) кожного окремого документу. Зрозуміло, що такий спосіб є практично неможливим, оскільки кількість документів, як правило, буває занадто великим, щоб всі їх можна було прочитати при кожному інформаційному запиті. Тому доводиться використовувати інший, менш ефективний метод, при якому ІП здійснюється не за самими текстами документів (умістом), а за короткими характеристиками змісту або певними зовнішніми ознаками документів. Для цього кожен документ забезпечується ***пошуковим образом документа*** (ПОД) – характеристикою, в якій стисло виражається основний зміст документу. Як було зазначено вище, інформаційний запит також має бути сформульований у вигляді такої ж короткої характеристики – ПОЗ. Завдяки цьому процедура ІП зводиться до зіставлення ПОД з заданим ПОЗ. Якщо ПОД з необхідною і достатньою мірою збігається з ПОЗ, вважається, що цей документ відповідає на інформаційний запит. Таке зіставлення виправдане лише тоді, коли пошуковий образ і пошуковий запит формулюються в термінах однієї мови, та ще такого, в якому кожна фраза допускає одне й тільки одне тлумачення.

ПОД містить загальний опис змісту документа. Тому такий метод не може забезпечити знаходження в бібліотеці всіх документів, які містять потрібну інформацію. Крім того, в масиві знайдених документів можуть бути такі, що фактично не відповідають даному інформаційному запиту. Такі документи створюють “пошуковий шум”.

Важливо пам’ятати, що інформація, яка міститься в наукових документах, об’єктивно підпорядковується закону розсіювання. Повнота і точність пошуку являють собою конкуруючі показники: підвищення одного з них веде до зниження іншого. При збільшенні повноти пошуку, ми неминуче зменшуємо його точність і, навпаки, збільшуючи точність пошуку, зменшуємо його повноту.

Ефективність інформаційного пошуку визначають показники, які характеризують знаходження релевантних документів. Вони підрозділяються на семантичні (*точність* та *повнота пошуку, коефіцієнт інформаційного шуму, коефіцієнт втрат* тощо) та техніко-економічні (оперативність пошуку, вартість та трудоємність пошуку).

Відповідність знайдених у процесі інформаційного пошуку знань або даних інформаційній потребі користувача (в особовому випадку – інформаційному запиту) називається ***пертинентністю***. Змістовна відповідність відображуваного результату його запиту за формальними (синтаксичними, морфологічними) ознаками називається ***релевантністю***.

З проблемою інформаційного пошуку першими зіткнулися бібліотекарі. Для того, щоб читачі могли знаходити в фондах бібліотеки документи, які їх цікавлять, в ній створювалися різні каталоги та вказівники. В одній з найбільших бібліотек давнини – в Александрійській бібліотеці – в 47 р. до н. е. нараховувалось біля 700 тис. томів (свитків папірусу). Складений Калімахом каталог до фондів цієї бібліотеки (приблизно в 250 р. до н. е.) мав обсяг 120 томів. Як основні елементи книгоопису в цьому каталозі використовувалися ім’я автора та назва (заголовок) твору. Якщо твір не мав назви, то Калімах приводив його початкові рядки.

Простішим ПОД є його заголовок. Спираючись на заголовок книги або статті читач у більшості випадків може судити про те, чи представляє для нього інтерес ця книга або стаття і чи варто з нею ознайомитися досконало.

Анотацію та реферат документу також можна вважати його пошуковими образами. Із збільшенням обсягу реферативних журналів кількість анотацій та рефератів, що містяться в них, стало настільки великим, що реферативні журнали довелося забезпечувати додатковим довідковим апаратом – системою покажчиків, які значно полегшують для читачів рішення інформаційно-пошукових задач. Таким чином, реферативні журнали, а також реферативні журнали з системою по-

кажчиків – це найпростіші документальні ІПС, розраховані на індивідуальне використання.

Існує три основних **типи інформаційно-пошукових задач**:

- ретроспективний інформаційний пошук, тобто пошук вже існуючих документів (всіх або частини), які містять відомості про певне питання;
- термінове сповіщення окремих спеціалістів (абонентів) про публікації, які мають для них потенційний інтерес. Даний тип інформаційного пошуку називається виборчим (адресним) розподілом інформації. Він виконується за постійними інформаційними запитами (так званими "профілями інтересів"), які формуються самими споживачами. Це окремий випадок інформаційного пошуку;
- пошук імен спеціалістів, які володіють інформацією з певного питання.

З розвитком Semantick Web та Web речей, цей перелік розширюється можливістю пошуку інформаційних об'єктів, які доступні за допомогою Інтернету.

2. Класична модель інформаційного пошуку

Базова стандартна модель, яка використовується в більшості книг з інформаційного пошуку виглядає, як показано на рис. 2 [5].

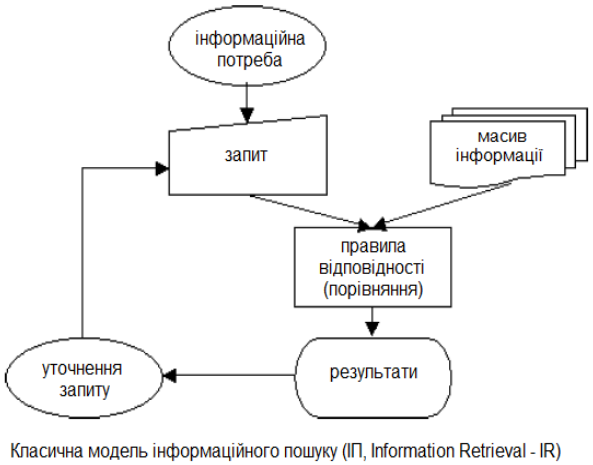

Рис. 2. Класична модель інформаційного пошуку

Як було зазначено вище, користувач, спонуканий інформаційною потребою, складає запит на деякій мові запитів. Запит посилається системі, яка вибирає з колекції документів (масив інформації) такі документи, що відповідають запиту згідно з визначеними правилами відповідності. Процес уточнення запиту може використовуватися для створення нових запитів та/або для очищення результатів.

Процес пошуку базується на використанні визначеної моделі пошуку. Модель пошуку характеризується наступними параметрами:

– форма подання документів і запитів;

– критерій змістовної відповідності;

– методи ранжування результатів запитів;

– механізм зворотнього зв'язку для оцінювання релевантності документів.

Наведемо стисло класичні моделі інформаційного пошуку:

- булева модель;
- ймовірнісна модель;
- векторна модель;
- дескрипторна модель та моделі, базовані на класифікаторах.

Булева модель. В цій моделі документ подається за допомогою набору термінів, які зберігаються в індексі. Кожен термін представлений як булева змінна. Документ (ПОД) подається як поєднання термінів. Вагові коефіцієнти не вводяться. Запит (ПОЗ) формується як довільний булевський вираз, що складається з термінів, пов'язаних логічними операціями (AND, OR, NOT). Мірою відповідності є значення статусу виборки (TRUE або FALSE). Така модель проста в реалізації і використовується в багатьох документальних ІПС. Ефективність пошуку невисока і неможливо ранжування документів за релевантністю.

Ймовірнісна модель. В основі ймовірнісних моделей лежить принцип його ранжування (Probabilistic Ranking Principle, PRP). Цей принцип заключається в наступному – найбільш загальна ефективність пошуку досягається у випадку, коли результативні документи ранжують-

ся за убуванням ймовірності їх релевантності запиту. Спочатку для кожного документу оцінюється ймовірність того, що він релевантний запиту, а потім за цими оцінками виконується ранжування документів.

Для отримання таких оцінок існують різні способи, а також додаткові допущення та гіпотези, які створені на основі апріорних відомостей про документи колекції. Відповідно до цього існує багато реалізацій ймовірнісної моделі пошуку. Наприклад, така оцінка може бути обчислена у відповідності з теоремою Байєса за деякою функцією ймовірностей входження термів даного документу в релевантні та нерелевантні документи. Використовуючи навчальну вибірку (навчальний масив даних) обчислюється ймовірність входження заданого терму в релевантні та нерелевантні документи [6].

Просторово-векторна модель (Vector Space Model) запропонована Солтоном в 1975 році, але на даний час має велике поширення. Векторні моделі, на відміну від булевих, дозволяють ранжувати результативну множину документів запиту. Документи (та запити до них) представляють собою набір векторів у n-мірному просторі [7]. Простір містить n базисних нормалізованих векторів, де n – загальна кількість різних термів в усіх документах. Значення компонентів вектора визначає вага терму (терміну). Показник відповідності (релевантності) визначається як оцінка кореляції між векторами. Така кореляція може бути скалярним добутком (множенням) вектора запиту на вектор документу [8]. Документи ранжують за спаданням скалярних добутків.

Дескрипторна модель є найпростішою моделлю пошуку. В ній документ задається у вигляді набору, асоційованих з ним зовнішніх атрибутів. У простих системах дескрипторного пошуку подання документу описується сукупністю слів або фраз лексики предметної області (ПрО), які характеризують зміст документа. Ці слова і словосполучення називаються дескрипторами. Індексування документу в таких системах реалізується призначенням для нього сукупності дескрипторів. При цьому дескриптори можуть приписуватися документу як на підставі його змісту, так і на підставі його назви. Такі два процеси називаються відповідно індексуванням документу за змістом та індексуванням за назвою [9]. В деяких дескриптивних системах індексування документів здійснюється вручну експертами ПрО, в інших воно виконується автоматично.

Дескрипторні системи можна віднести до класу систем, орієнтованих на бібліографічний пошук або пошук у каталозі.

Моделі, базовані на класифікаторах – є однією з різновидів найпростіших моделей пошуку. Документ у цій моделі, як і в дескриптивних системах, подається у вигляді сукупності асоційованих з ним атрибутів. Атрибутами є ідентифікатори класів, до яких відноситься даний документ. Класи формують ієрархічну структуру класифікатора. Запит може бути представлений двома способами:

– простий варіант, коли запитом є ідентифікатор будь-якого класу з заданого класифікатора. Критерій релевантності документу запиту – клас документу збігається з класом, поданим у запиті, або є його підкласом;

– складний варіант – в запиті можна вказати кілька класів класифікатора. Критерій релевантності документу запиту – клас документу збігається з будь-яким із зазначених у запиті класом, або є його підкласом.

Моделі, базовані на класифікаторах, близькі до булевських моделей.

Необхідно зазначити, що класичні моделі розглядають незалежність слів (термів). Для подання документів та запитів застосовується одразу декілька моделей.

Ефективність пошуку (інформаційно-пошукових систем) аналізується і регулюється перш за все за рівнем релевантності й пертинентності в частині вдосконалення організації запитів користувачів, пошуку за параметрами, за рахунок кластеризації, пошуку за подобою, ранжуванням відгуків, використання «сюжетних підходів», всебічного використання семантичних методів (у тому числі із застосу-

ванням автоматичного групування документів за класифікатором, автоматичним визначенням раніше незаданих або слабо структурованих документів, ранжування документів за змістовою релевантністю, автоматичного аналізу та змістовного перетворення запитів, виявлення семантично подібних документів на зразок порівнянню з еталоном, наприклад, з використанням матриці Александера).

3. Типи пошуку

Інформаційний пошук можна розділити на наступні види:

- ***повнотекстовий пошук*** – при цьому здійснюється пошук в усьому змісті документу. Прикладами повнотекстового пошуку є більшість пошукових систем Інтернету, як Yandex, Google тощо. Зазвичай, для прискорення пошуку повнотекстовий пошук використовує попередньо створені індекси (індексну базу);
- ***пошук за метаданими*** – це пошук за деякими атрибутами документу, які підтримуються системою. Наприклад, назва документу, дата створення, розмір, автор тощо. Прикладом пошуку за реквізитами є діалог пошуку в файловій системі (наприклад, в ОС MS Windows). Цей пошук зазвичай використовує дескриптивну модель пошуку;
- ***пошук зображення*** – це пошук за вмістом зображення. Пошукова система зазвичай використовує алгоритми штучного інтелекту – порівняння за зразком та пошуку за подібністю;
- ***пошук музики*** – аналогічно пошуку зображення, виконує пошук за зразком у колекції музичних даних;
- ***пошук інформаційних об'єктів*** здійснюється в середовищі Web речей; виконує комбінований пошук інформаційних об`єктів, що доступні в Інтернет, з використанням мета-описів цих об'єктів та з урахуванням типу об'єкта.

4. Класифікація видів пошуку

Адресний пошук. Процес пошуку документів здійснюється за суто формальними ознаками, які вказані у запиті. Для здійснення такого типу пошуку необхідні наступні умови:

– наявність у документі точної адреси;

– забезпечення суворого порядку розташування документів у запам'ятовуючому пристрої або в сховищі системи.

Адресами документів можуть бути адреси Web-серверів та Web-сторінки, елементи бібліографічного запису, адреси зберігання документів у сховищі.

Документальний пошук. Процес пошуку здійснюється в сховищі інформаційно-пошукової системи первинних документів або в базі даних вторинних документів, що відповідають запиту користувача.

Існує два різновиди документального пошуку:

– бібліотечний, який спрямований на знаходження первинних документів;

– бібліографічний, який спрямований на знаходження відомостей про документи, які подані в вигляді бібліографічних записів.

Фактографічний пошук. Процес пошуку полягає у пошуку фактів, які відповідають інформаційному запиту. До фактографічних даних відносяться відомості, які добуті з первинних або вторинних документів, або які отримані безпосередньо з джерел їх виникнення.

Розрізняють два підвиди фактографічного пошуку:

– документально-фактографічний, який полягає у пошуку в документах фрагментів тексту, які містять факти;

– фактологічний (опис фактів), який припускає створення нових фактографічних описів у процесі пошуку шляхом логічної обробки знайденої фактографічної інформації.

Семантичний пошук. Цей пошук полягає у пошуку документів за їх змістом. Для здійснення такого типу пошуку необхідні наступні умови:

– переклад змісту документів і запитів з природної мови на інформаційно-пошукову мову для створення пошукових образів документу і запиту;

– створення пошукового опису, в якому вказується додаткова умова пошуку.

Принципова різниця між адресним та семантичним пошуками полягає у тому, що при адресному пошуку документ розглядається як об'єкт з точки зору форми, а при семантичному пошуку – з точки зору змісту. При семантичному пошуку знаходиться множина документів без зазначення адрес. Це є принциповою відмінністю каталогів і картотек. Бібліотека – це є зібрання бібліографічних записів без вказування адрес.

5. Інформаційний пошук у Web-середовищі

Поява та розвиток Інтернету сприяли розширенню поняття пошуку та появі більш специфічного поняття Web-пошуку. Оскільки в контексті Web фактори взаємодії людини з комп'ютером та когнітивні аспекти грають найважливішу роль, корисно деталізувати цю модель, як показано на рис. 3.

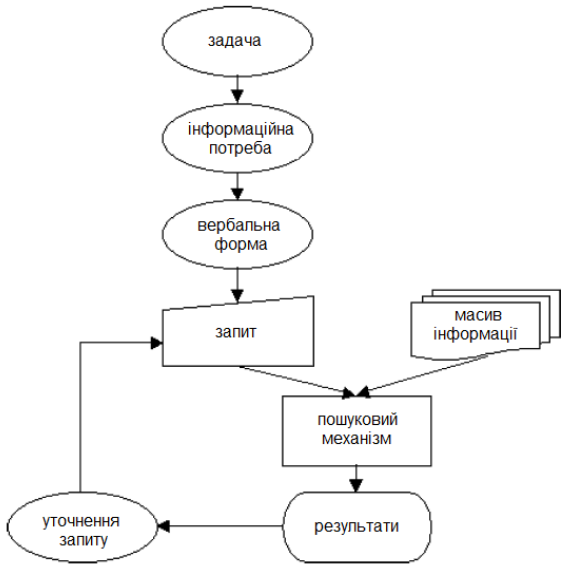

Рис. 3. Класична модель інформаційного пошуку, поширена на Інтернет-мережу (Web)

Як було зазначено раніше, інформаційна потреба асоціюється з (викликається) деякою задачею. Ця потреба вербалізується (найбільш часто це виконується ментально та не дуже чітко) та транслюється в запит, що надається пошуковому механізму. Цей процес висвітлення та створення запиту з інформаційної потреби, в контексті Web здобув велику увагу: в статті Хольстера та Струбе [10] вказується на тому, що досвідчені користувачі та новачки конструюють запити по-різному. Навarро – Пьєтро та ін. [11] вивели когнітивну модель для Web-пошуку, Мурамату та Прат [12] дослідили ментальну модель користувачів пошукових механізмів тощо. Також у [13] необхідно зауважити, що всі ці дослідження базуються на припущенні, що Web-пошуковці мотивовані (спонукувані) інформаційною потребою.

5.1. Таксономія Web-пошуку. В контексті Web, вираз "потреба спонукає запит" часто не є інформативним. У 2002 році автор [14] класифікував запити у відповідності до їх намірів на три наступних класи:

- ***навігаційні запити***. Такі запити мають на меті негайний намір побачити певний сайт;
- ***інформаційні запити***. Вони виражають намір отримати деяку інформацію, яка вважається існуючою на одній або більше Web-сторінках;
- ***транзакційні запити***. Ці запити виражають намір виконати якусь Web-опосередковану діяльність – покупку в Інтернет-магазині, завантаження файлів тощо.

Навігаційні запити. Метою таких запитів є дістатися певного сайту, який користувач має на увазі. Це визначено тим, що користувач можливо відвідував цей сайт у минулому, або він припускає, що такий сайт існує. Наприклад:

Запит	Можливий результат
compaq	Http://www.compaq.com
Фуршет	http://www.furshet.ua/
Газета по-киевски	http://mycityua.com

Цей тип пошуку іноді вважається, як пошук "загальновідомого предмету" в класичному ІП. Прикладом такого пошуку

стало завдання "Пошук домашньої Web-сторінки", яке регулярно проводиться при тестуванні пошукових систем при конференції з текстового пошуку (Text Retrieval Conference).

Навігаційні запити зазвичай мають тільки один правильний результат.

Транзакційні запити. Мета таких запитів полягає у тому, щоб досягти місця (сайту), де можливо провести подальшу взаємодію (транзакція) для досягнення певної мети. До основних категорій для таких запитів можна віднести здійснення покупок, пошук різних Web-опосередкованих сервісів, завантаження різного типу файлів (зображень, пісень і т. д.), доступ до деяких баз даних (наприклад, типу Yellow Pages), пошук серверів (наприклад, для ігор) і т. д.

Результати таких запитів з точки зору класичного ІП дуже важко оцінити. Все, що можливо – це бінарне значення оцінки, скажімо, відповідно чи не відповідно. Проте найбільш важливі для користувачів зовнішні чинники (наприклад, ціна товару, швидкість обслуговування, якість і таке інше), як правило, в загальних пошукових системах недоступні.

Інформаційні запити. Метою таких запитів є знайти інформацію, яка припускається існує у Webі в статичній формі. В подальшому взаємодій ніяких не передбачається, за винятком читання. Під статичною формою мається на увазі, що цільовий документ не створюється як відповідь на запит користувача. Ця різниця дещо розмита, оскільки змішування результатів, що характерно для третього покоління пошукових систем, можливо, призведе до використання динамічних сторінок.

В будь-якому випадку, інформаційні запити – найбільш приближені до класичного поняття інформаційного пошуку (IR), і тому вони далі будуть розглянуті детальніше.

На відміну від звичайного пошуку, більшість інформаційних запитів, що здійснюються в Інтернеті, семантично є надзвичайно широкими, наприклад, "автомобілі" або "Сан-Франциско", водночас як деякі можуть бути вузькими, наприклад "normocytic anemia" або "метрична система". Досліди інформаційних запитів, проведені в [14] відзначають, що майже 15 % усіх пошуків за бажану мету вважають гарну колекцію посилань за заданою темою, ніж один добрий документ.

Експериментальні результати дослідження типів запитів надані в таблиці.

Таблиця. Класифікація запитів користувачів

Type of query	User Survey	Query Log Analysis
Navigational	24.5 %	20 %
Informational	?? (estimated 39 %)	48 %
Transactional	> 22 % (estimated 36 %)	30 %

Пошукові системи необхідні для вирішення всіх трьох типів запитів, хоча кожен тип задовольняється досить різними результати. Розуміння цієї таксономії має важливе значення для успішного розвитку Web-пошуку. Сучасні пошукові системи добре вирішують інформаційні та навігаційні запити, але транзакційні запити задовольняються лише опосередковано. Шлях підвищення ефективності пошуку лежить в удосконаленні семантичного аналізу (тобто розуміння того, про що запит) та змішування різних зовнішніх баз даних.

5.2. Визначення пошуку в Web-середовищі. В зв'язку з появою Web, поняття пошуку в середовищі Інтернету набуло іншого змісту. Поняття пошукової системи стало більш широким та глибшим. Наведемо декілька новітніх визначень поняття пошукової системи (Search Engine), що прийняті нині в західній науковій літературі.

Пошукова система – це комп'ютерна програма, яка отримує (retrieves) файли або документи, або дані з бази даних або з комп'ютерної мережі (зокрема, з Інтернету) [15].

Пошукова система – це комп'ютерна програма, яка знаходить (finds) інформацію в Інтернеті шляхом пошуку слів,

які були введені (як запит – уточнення автора) [16].

Пошукова система – це комп'ютерне програмне забезпечення для пошуку даних (з текстів або баз даних) для отримання конкретної інформації, а також: сайт у Web-мережі, який використовує програмне забезпечення для пошуку ключових слів на інших сайтів [17].

В контексті Web з огляду на тезу, що „потреба спонукає запит", у клас поняття пошукових систем почали включати системи „запитання-відповідь" (answer engine), які дуже часто є фактографічними ПС. Але деякі системи для отримання результату пошуку вже починають використовувати процедури логічного виводу.

Зважаючи на вищесказане, пошукова система, в контексті Web, використовує спеціалізоване програмне забезпечення, яке має на вході від користувача пошуковий/і термін/и і на виході надає список Web-сторінок, які вважаються найбільш релевантними. Більшість пошукових систем мають величезні бази даних мільярдів Web-сторінок. Розрізняють два типи Web-пошукових систем: пошукові системи, базовані на кроулінгу та каталоги.

Пошукові системи, базовані на кроулінгу (Crawler-based). Такі системи створюють свої списки Web-сторінок автоматично. Вони "сканують" (crawl) Інтернет за допомогою робота-"павука" (spider, програма, яка відвідує Web-сторінки, читає їх і слідує далі за посиланнями, знайденими на Web-сторінці), і повертають користувачу результати пошуку, які ранжовані у порядку важливості. Павук повторно відвідує Web-сторінки кожні кілька місяців для найчастішого оновлення своєї індексної бази відповідно до внесених на Web-сторінки змін. Головна перевага пошукових систем, базованих на кроулінгу, полягає у тому, що будь-які зміни, які внесені до Web-сторінки, будуть впливати на його базу і відповідно – результати пошуку. Таким чином, актуальність змісту Web-сторінок збігається з ключовими словами, що використовуються для пошуку.

Каталоги, що створені людиною (human based directory), залежать від людей, які його створили та поповнюють. Вони виконують пошук за ключовими словами в коротких описах Web-сторінок, представлених Web-майстрами та спеціалістами, що рецензують та перевіряють каталог. Разом з цим, Web-сторінки переглядаються людиною і розміщуються в відповідну ієрархію категорій. Таким чином, зміни, внесені до Web-сторінки, на відміну від скануючих пошукових систем, не будуть мати ніякого впливу на збережений у каталозі опис. Отже, хоча на Web-сторінці і міститься відповідна інформація, яка відповідає запиту, але вона не буде відображена в списку результатів пошуку доки Web-майстер не змінить опис Web-сторінки. Саме з цієї причини один з найперших та найбільших каталог, сформований людиною Yahoo! перетворено у більш популярну пошукову систему на базі сканеру. Таким чином утворюються комбіновані пошукові системи. Оскільки каталоги містять інформацію, перевірену людиною, ця інформація використовується для фільтрування та ранжування результатів пошуку.

Окрім зазначеного вище, розрізняють наступні **типи пошукових механізмів**:

- пошукові системи;
- Web-каталоги;
- віртуальні бібліотеки;
- мета-пошукові механізми.

Пошукові системи (Search Engines) є найбільш широким класом ІПС та найбільш популярним і загальновживаним. Вони характеризуються наступними властивостями:

- мають базу даних Web-сторінок;
- пошук здійснюють за ключовими словами;
- мають скануючого робота.

Яскравим прикладом такої системи є пошукова система Google.

Web-каталоги (Web Directories). Як було вказано вище, вони:

- мають колекцію Web-ресурсів;
- організовані за тематичними категоріями в ієрархію;

- організація в категорії та інше проводиться вручну.

Приклад такого каталогу – загальновідомий каталог Yahoo.

Віртуальні бібліотеки (Virtual Libraries). Такі бібліотеки характеризуються наступними ознаками:

- мають колекцію Web-джерел;
- оцінюються фахівцями з предметної області;
- слабо автоматизовані, живляться людськими ресурсами.

Приклад типової бібліотеки – бібліотечний індекс Інтернету – Librarians Index to the Internet www.lii.org.

Мета-пошукові механізми (Meta-Search Tools). З назви видно, що такі механізми використовують ресурси інших пошукових систем, а результати фільтрують та ранжують згідно своїх заданих правил. Такі системи характеризуються:

- не мають власної бази даних;
- вони здійснюють запити до інших пошукових механізмів, розташованих у Web;
- мають дуже поганий дизайн і можуть тільки змінювати порядок ранжування результатів.

Класичний приклад такої системи є MetaCrawler.com. Такі системи користуються попитом, оскільки вони повертають більш короткий список посилань, що психологічно більш прийнятно для людини.

5.3. Еволюція пошукових систем інтернет. У зв'язку з таксономією, наведеною вище, в 2002 році в [14] було визначено три етапи (генерації) розвитку Web-пошукових систем.

Перше покоління пошукових систем використовувало в основному інформацію, яка знаходилась безпосередньо на Web-сторінках (текст і форматування), ці пошукові системи дуже близькі до класичних ПС. Такі системи виконують в основному тільки інформаційні запити. Типовими прикладами таких систем в 1995–1997 роках були загальновідомі AltaVista, Excite, Webcrawler і т. д. Ранжування сайтів відбувалося тільки за рахунок контенту сторінок.

Важливі фактори, які враховувалися при ранжуванні, включали щільність ключових слів на Web-сторінці, назву, і місцезнаходження цих ключових слів у цьому документі. Також ПС першого покоління для обчислення релевантності враховували мета-тегі, використання ключових слів в імені домену, а також в URL-адресі (докладніше – див. [29]).

Основні спам-фільтри робили перевірку на наявність ключових слів у тексті, представлених на сторінці тим самим кольором, що і фон документу, тобто невидимих людському зору. На той час з'явилися перші портали, в наслідок чого результати пошуку перетворилися у величезні рекламні щити та перевантажені інформацією жовті сторінки.

Друге покоління пошукових систем (початок появи 1998–1999 рр.) характеризується використанням інформації, яка існує поза Web-сторінкою – Web-специфічних даних таких, як аналіз посилань (link analysis), тексту якорів (anchor-text) та відстеження даних, що передаються з http-запитом (click-through data). Таким чином вони стали брати до уваги структуру Web-мережі.

Друге покоління більш щільно пов'язано з семантикою запитів, яка береться з аналізу даних, що подані у Webі поза сторінки. Деякі з основних компонентів, які вони використовують є відстеження кліків (tracking clicks), репутація сторінки (page reputation), індекс популярності (link popularity), темпоральні спостереження (temporal tracking, кількість часу, що проводять відвідувачі на сторінці), та якість посилань (link quality). Пізніше, ПС другого покоління почали використовувати вектори термів (term vectors) [18], аналіз статистики відвідування (stats analysis), кеш-дані (cache data) і контекст. Як аналіз контексту розглядається пошук на сторінці пар ключових слів, які складаються з двох слів. Це дозволяє краще виконати віднесення сторінки до певної категорії.

Першою системою, яка почала використовувати аналіз посилань між сторі-

нками як один з основних факторів ранжирування, стала система Google (PageRank). ПС DirectHit стала першою, хто побудував ранжування на аналізі даних, що передаються під час http-запиту. В даний час всі основні системи використовують всі ці типи даних. Використання Google PageRank та метод відстеження кліків DirectHit та тривалості візиту, підвищило ефективність пошуку.

Пошукові системи другого покоління підтримують як інформаційні, так і навігаційні запити. Аналіз посилань та текст якорів мають вирішальне значення для навігаційних запитів.

Третє покоління пошукових систем. На даний час третє покоління пошукових систем знаходиться в стані зародження та початкового розвитку. Ці пошукові системи є спробою поєднати дані з різних джерел для досягнення головної мети – видачі результату, що відповідає потребі користувача. Наприклад, на запит „Ялта", ПС має надавати пряме посилання на сторінку бронювання готелів у Ялті, сервер мап з мапою міста, на сервер погоди з інформацією про погоду і т. д. Таким чином, третє покоління – це покоління пошукових систем, які виходять за рамки обмежень фіксованої бази даних за допомогою семантичного аналізу, визначення контексту пошуку, вибору динамічної бази даних і т. д. Завдання полягає у тому, щоб забезпечити інформаційні, навігаційні і транзакційні запити.

Третє покоління пошукових технологій покликані об'єднати масштабованість існуючих Інтернет-пошукових систем з новими та удосконаленими моделями пошуку релевантності; вони починають враховувати вподобання користувача, співробітництво, колективний інтелект, багатий досвід користувачів, та багато інших спеціалізованих можливостей, які роблять інформацію більш значимою, а пошук – більш продуктивним.

Пошукові системи третього покоління додають до бази даних векторів термів похідні слова (word stemming) і тезаурус, що надає допомогу у здійсненні пошуку за контекстом [19]. Автоматичне визначення ключових пар також допомагає автоматичній категоризації сторінки, визначенню де користувач хоче провести покупку, а де – здійснити пошук, що має видати абсолютно різні результати пошуку на основі контексту або намірів користувача.

Технології третього покоління збагачені картами Web, які є корисними для фільтрації – видалення дублікатів сайтів, а також багатьох самостійних сторінок, які привертають трафік на всього лише декілька ключових слів. Це означає, що сторінки типу дорвеєв (doorways), гейтвеєв (gateways), вхідних (entry, splash) – спеціально створені спам-сторінки для цільової розкрутки сайту на визначені позиції ключових слів, незабаром будуть відфільтровані.

Вони також будуть витягувати як можна більше даних про індивідуальні пошукові звички користувача. Всі основні пошукові системи планують створення персональних профілів та агентів, які будуть накопичувати знання про користувача протягом певного періоду часу та використовувати їх виходячи з минулих пошукових звичок.

Поява Семантичного Web (докладніше див. [20]) надало нові можливості і ще більше диференціювало поняття інформаційного пошуку. Семантичний Web надав можливість використовувати існуючу семантичну інформацію – подану за допомогою семантичної розмітки, використовуючи семантичні зв'язки, виконуючі різні операції виведення на семантичних даних, а також порівняння семантичної інформації. Змінюється і алгоритм ранжування результуючих документів – вводиться поняття семантичного ранжування документів. Змінюється алгоритм пошуку, він стає дедалі розподіленим, змінюються методи задання пошукового запиту. Поява різних типів поданої у Web інформації (різної модальності – мультимедійної інформації, відео, аудіо тощо) потребує використання інших підходів. Існуюче розділення пошуку за типом інформації – пошук відео, пошук картинок, тощо (Google, Яндекс) – дуже стиснено і неінформативне. Існує синергетична потреба – виконання пошуку в різних

типах інформації та подальше змішування результатів.

Поява нового явища – Web речей (Web Of Things), який містить не тільки звичні документи, але й електронні пристрої та інші побутові речі, які підключені до Інтернету і можуть керуватися і знаходитися віддалено, також потребує врахування таких нових типів інформаційних об'єктів.

Таким чином, пошукові системи 3-го покоління виходять за рамки класичного (традиційного) поняття пошуку в зв'язку з появою нових типів інформації та нових вимог, що ставлять користувачі перед пошуковими системами.

В західній літературі з`явився термін Search 2.0, який асоціюється з третім поколінням, але має більш чіткі обриси і більш орієнтовано на бізнес-аудиторію [21]. У Webі вже існує десяток проектів, які вважаються проектами Search 2.0 – Swicki (http://www.swicki.com/), Rollyo (http://www.rollyo.com/),
Clusty (http://www.clusty.com/),
Wink (http://www.wink.com/),
Lexxe (http://www.lexxe.com/), тощо.

5.4. Приклади технологічних рішень пошукових систем третього покоління. З розвитком нових технологій та стандартів, паралельно з науковими дослідженнями, та спираючись на них, компанії бізнес-сектору прагматично розвивають нове покоління пошукових систем – «розумних» ПС, "smarter" search engines. Наведемо приклади таких технологічних рішень пошукових систем, які інтелектуалізують процес пошуку за рахунок:

- структурування та представлення (подання) даних, отриманих з Інтернету;
- реалізації семантичної фільтрації за якістю;
- організації пошуку серед структурованих даних в Інтернеті;
- пошуку в режимі реального часу в Інтернеті;
- пошуку в «глибинному» Web ('deep web') [22].

Структурування та подання даних

Wolfram Alpha (Система обчислювання знань, Computational Knowledge Engine, http://www.wolframalpha.com/, 2009). Цей амбіційний проект стартував 5 березня 2009 року. Автор цього Web-сервісу – британський фізик Стівен Вольфрам (Stephen Wolfram), голова компанії Wolfram Research, розробник широко відомої у наукових колах програми Mathematica.

На відміну від традиційних пошукових систем, які обмежуються тим, що за запитом користувача видають список посилань на сайти, які мають відповідати запиту, сервіс Wolfram Alpha самостійно аналізує запити користувача і представляє йому зведену релевантну інформацію.

З огляду на прийняту класифікацію ця система є системою „питання-відповідь". Автор позиціонує систему не як пошуковий сервіс (search engine), а як Computational Knowledge Engine («систему обчислювання знання»).

Ця система об'єднує обчислювальні потужності Mathematica з інструментами, які експліцитно оперують з усіма типами даних з тим щоб надати точну відповідь на запитання, яке сформульоване в природньомовній формі, в будь-яких можливих предметних областях [23]. Оскільки ця система є бізнес-застосуванням, докладного опису її функціювання у вільному доступі не має.

Спочатку Wolfram Alpha працював у закритому (тестовому) режимі, а з 18 травня 2009 р. Web-сервіс відкритий для всіх бажаючих. За час закритого тестування було оброблено близько 23 млн. запитів, а за перший тиждень після відкриття – близько 100 млн. На сьогоднішній день Wolfram Alpha є безкоштовним Web-сервісом.

Предметні області, які обробляються в системі – математика, фізика, хімія, астрономія, статистика та дані статистичного аналізу, дати та час, географія, погода, здоров'я та медицина, культура та медіа, музика та освіта, люди та історія, фінанси, лінгвістика і досягнення високих технологій, спорт тощо.

Можливості системи [24]:

- переведення одиниці виміру з однієї системи в іншу;
- якщо задати хімічну формулу, система видасть основну інформацію про цю речовину / хімічний елемент;
- якщо ввести в рядок пошуку 1 apple + 1 orange, – система видасть кількість калорій, протеїнів, вітамінів, відсутність / наявність холестерину і т. д.;
- якщо ввести назву міста, то система видає інформацію про те, де воно знаходиться, кількість жителів, схематичне розташування на карті, поточний час, поточну температуру, вологість, швидкість вітру, стан хмарності, висоту над рівнем моря, найближчі міста (з відстанню до них і з кількістю мешканців у цих містах). Натиснувши на посилання „Show coordinates", можна дізнатися координати міста. Натиснувши на посилання „Satellite image", система завантажить знімки міста з супутника (буде завантажений сайт "Карти Google");
- система виконує різні обчислення: якщо ввести в рядок пошуку, наприклад, $ 999 + 15 %, Wolfram Alpha зробить необхідні обчислення;
- система надає інформацію про будь-який сайт. Якщо ввести в рядок пошуку URL сайту, система видасть детальну інформацію: хто є хостинг-провайдером, де він розташований, кількість переглядів і кількість візитів за добу, site rank, найменування і розмір титульної сторінки, кількість вихідних посилань, кількість «зображень»;
- система може проводити не тільки найпростіші обчислення, але й вирішувати різні рівняння: якщо ввести, наприклад, x ^ 3 sin (x), система видасть рішення у вигляді графіка та в аналітичному вигляді;
- обробка музики, якщо ввести в рядок пошуку, наприклад, C Eb GC, то система надасть вичерпну інформацію про ці музичні ноти;
- обробка імен, якщо ввести два різних імені, наприклад, Vera, Natasha, в результаті система видає статистичні дані, що свідчать про те, як часто використовуються ці імена;
- обробка фінансової інформації: система може надавати інформацію про економічний стан (наприклад, про наявність акціонерного капіталу, вартості однієї акції і т. д.) двох компаній, назви яких вводяться у пошуковий рядок з пробілом між назвами;
- обробка часової інформації: якщо ввести дату в форматі, наприклад, august 28, 1959, то система видасть, який це був день тижня, можна буде підрахувати, скільки часу (років, місяців, тижнів, днів) пройшло з цієї дати, хто з відомих людей народився в цей день, які свята припадають на цей день.

Для того, щоб дізнатися джерела інформації, які використовував Wolfram Alpha, унизу, під знайденої інформацією знаходиться кнопка „Sources".

Всю інформацію, яку згенерував («навольфраміл» – сленг) Wolfram Alpha, можна зберегти у вигляді PDF-файлу.

Нажаль, система обробляє тільки англомовні запити.

Google Squared

Google Squared – цей експериментальний пошуковий механізм було заявлено 3 червня 2009 р. На відміну від класичних – «традиційних» пошукових систем, Google Squared не видає на запит користувача сторінку зі списком посилань на Web-ресурси, що відповідають запиту. Як результати пошуку користувачу виводиться зведена таблиця з інформацією з запросу. Тобто Google Squared, як і сервіс Wolfram Alpha, самостійно аналізує (намагається аналізувати) запити користувача і надає йому зведену релевантну інформацію.

В офіційному блозі пошукового гіганта сказано так: «...Squared Google не шукає Web-сторінки за вашим запитом…він автоматично вибирає і організовує факти зі всього Інтернету» [25].

Як і Wolfram Alpha, сервіс Google Squared не підтримує українську та російську мови.

Порівняльне тестування Google Squared та Wolfram Alpha, наведене авто-

ром у червні 2009 р. в [26] показує, що аналітичні характеристики і можливості системи Google Squared на даний час явно поступаються Wolfram Alpha.

Google Squared був експериментальним проектом, в якому корпорація Google проводила тестування функціоналу роботи пошукової системи з урахуванням структурованої інформації та початком інтелектуальної обробки знань. На даний час проект закрито.

Sensebot

SenseBot (http://www.sensebot.net/, 2008 р.) заявлена як семантична пошукова система, яка на пошуковий запит генерує текстові анотації (резюме), складені з Web-сторінок, які відносяться до теми пошукового запиту. Ця система для вилучення змісту з Web-сторінок і представлення його користувачеві узгодженим чином використовує інтелектуальну обробку текстів (text mining) і мультидокументну сумарізацію (multidocument summarization). Разом з результатами система видає „семантичну хмару" концептів ("Semantic Cloud" of concepts), що дозволяє направити увагу та керувати результатами.

Оскільки SenseBot є семантичною пошуковою системою, це означає, що вона намагається зрозуміти семантику отриманих у результаті сторінок. Вона використовує, як було зазначено вище, інтелектуальну обробку текстів для розбору Web-сторінок і визначення їх основних семантичних концептів.

На верхньому рівні, система отримує джерела, які видаються пошуковою системою як результат. Після цього система виконує інтелектуальну обробку тексту, отриманого з кожного джерела, вилучаючи ключові концепти. Подібності між джерелами оцінюються і ті, що семантично знаходяться далеко від запиту або не зв`язані з загальною масою знайдених джерел, відкидаються. Концептам присвоюється вага, а також для концептів, які представлені у запиті, задаються преференційні значення. Після чого виконується відповідно до запатентованого алгоритму мультидокументна сумаризація – збір підсумкового документу, складеного з текстів резюме, які згенеровані зі знайдених документів. Таким чином, на запит користувача фактичними результатами Web-пошуку є резюме, згенероване зі знайдених документів.

Найкращі результати можуть бути досягнуті на множині текстових документів, які по суті знаходяться близько до заданої теми. Найкраща область застосування цієї системи, як зазначає її розробник, є вертикальні пошукові системи і портали – фінансові, медичні, правові, бібліотеки і т. д. Що стосується загального Web-пошуку, деяка кількість "шуму" неминуча, навіть для тих джерел, що знаходяться на перших сторінках результатів, які вважаються найбільш релевантними [27].

Реалізація семантичної фільтрації інформації за якістю

Hakia

Цей відомий проект (http://www.hakia.com/) засновано в 2004 р. Для роботи системи була розроблена альтернативна інфраструктура, яка використовує алгоритм SemanticRank, який використовує онтологічну семантику, обчислювальну лінгвістику та нечітку логіку. На час, коли система була в відкритому доступі, вона охоплювала тільки предметну область з медицини та здоров'я. Заявлялося, що семантична технологія Hakia забезпечує новий досвід пошуку, який орієнтований на якість, а не популярність. Для проведення подальшого дослідження в галузі ІП досить корисними є основні 3 критерії, яким одночасно мають задовольняти якісні результати:

– якісні результати надходять з заслуговуючих довіри Web-сайтів, рекомендованих бібліотекарами або довіреними особами;

– якісні результати представляють собою найбільш свіжу наявну інформацію;

– якісні результати залишаються абсолютно релевантними до запиту.

Проект був відкритий для користування до квітня 2014 р. На даний час його повністю закрили і надалі використовують

для закритих комерційних рішень з підтримки Web-сайтів з обмеженою ПрО.

Організація пошуку серед структурованих даних у Webі

SWSE

На даний час вже існує багато даних, які відповідають запропонованим стандартам Семантичного Weby (наприклад RDF та OWL). Вже існує багато малих вертикальних словників і онтологій, які все більше використовуються різними спільнотами для вирішення своїх конкретних задач: користувачі Weby публікують описи своїх профілів з використанням формату FOAF (Friend of a Friend), провайдери новин транслюють добірку новин у вигляді RSS (RDF Site Summary), зображення анотуються з використанням різноманітних RDF-словників тощо.

SWSE (http://swse.deri.org/) представляє собою сервіс, який постійно вивчає та індексує Семантичний Web (Semantic Web) і забезпечує легкий у використанні інтерфейс, за допомогою якого користувачі можуть знайти дані, які вони шукають.

SWSE індексує триплети RDF або OWL, знайдені в Web, і надає послугу з пошуку серед цих триплетів.

На даний час проект закритий для зовнішнього використання і інтегрований у загальні проекти консорціуму W3C.

Swoogle

Swoogle (http://swoogle.umbc.edu/) –пошукова система, створена спеціально в рамках розвитку Семантичного Web. Кроулери Swoogle сканують Web з метою пошуку спеціального класу Web-документів, які називаються семантичними Web-документами, тобто які написані мовами RDF або OWL. Ця пошукова система індексує знайдені семантичні документи і зберігає їх, поступово формуючи онтологічну базу знань, та виконує пошук серед RDF-триплетів, видаючи в результатах пошуку посилання на джерела, які їх містять та фрагменти відповідних онтологій. Пошук здійснюється за ключовими словами та з використанням додаткових онтологічних конструкцій – обмежень.

Аналогічні функції пропонують і пошукові системи WatsOn, Semanticweb-search, Sindice (http://sindice.com/), Falcons.

Пошук у Web у режимі реального часу

Topsy

Пошукова система **Topsy** (http://www.topsy.com/) у режимі реального часу сканує інформацію, яка постійно генерується користувачами соціальних мереж Twitter, Digg, тощо. Якщо повідомлення містить посилання на Web-сторінку, то в разі, якщо за алгоритмом системи таке посилання буде вважатися важливим, воно буде проіндексоване. Таке індексування пошукова система проводить у режимі реального часу – поява нового посилання на сервісі одразу викликає процес індексування. Алгоритм визначення важливості посилань враховує багато умов, одними з головних є авторитетність джерела інформації та рівень довіри (trust).

Кінцевим результатом роботи пошукової системи є пошуковий досвід, який дозволяє користувачам знаходити свіжий, найбільш соціально значущий контент у реальному часі у Web. Результати пошуку індексуються в залежності від їх актуальності та популярності. Окрім текстової інформації, система індексує фото та відео, а також інформацію з соціальних мереж (твіти, пости тощо).

Scoopler

Scoopler – один з найкращих стартапів, який запропонував виконання пошуку в режимі реального часу. Аналогічно пошуковій системі Topsy, робот цієї ПС збирає і організовує контент актуального типу – новини, фотографії та відеоматеріали значних подій, а також посилання на найгарячіші нотатки поточного дня. Джерелами контенту, який індексується, є постійні оновлення, що поступають з сервісів Twitter, Flickr, Digg, Delicious тощо. На даний час система викуплена корпорацією Google і використовується для пошуку в соціальній мережі Google+.

Пошук в «глибинному» Web ('deep web')

DeepDyve

DeepDyve(http://www.deepdyve.com/) – пошуково-„дослідницька" система, яка використовує власні (комерційні) технології пошуку та індексування, що дозволяють відбирати багатий, релевантний контент з тисяч журналів, мільйонів документів і мільярдів незадіяних Web-сторінок глибинного Web. Дослідники, студенти, технічні спеціалісти, бізнес-користувачі, а також споживачі іншої інформації, можуть отримати доступ до багатої інформації, що зберігається в „глибинному Web" – інформації, яка складає переважну більшість в Інтернеті, але не індексується традиційними пошуковими системами. Пошуково-дослідницька система DeepDyve відчиняє шлях до цього поглибленого професійного контенту і повертає результати, які не навантажені інформацією з оглядових (реферативних) сайтів та іншою нерелевантною інформацією.

Система використовує запатентований алгоритм KeyPhrase ™, який застосовує метод індексації, отриманий при дослідженнях в області геноміки. Алгоритм шукає збіг патернів і символи за спеціальною метрикою. Система знаходить відповідність документів там, де традиційні пошукові системи нічого не знаходять. Тому ця система ідеально підходить для пошуку складних даних, що містяться в глибинному Web.

Також існує багато пошукових систем, що виконують пошук у глибинному Web, які спеціалізуються на конкретній предметній області та містять перевірені і рецензовані спеціалістами статті. Такі ПС, як правило, мають вузько спрямовані репозиторії, що надає реальну перевагу для цілеспрямованого пошуку дослідника в певній ПрО.

До таких спеціалізованих порталів можна віднести Mednar (www.mednar.com) – портал з глибинного пошуку в галузі медицини, Biznar (www.biznar.com) – пошук в бізнес-галузі, Worldwidescience (www.worldwidescience.org) – глобальний науковий портал, Science.gov (www.science.gov) – науковий портал уряду США, Scitopia (www.scitopia.org) – пошукова система наукової інформації і патентів, Nutrition.gov (www.nutrition.gov) – портал, який містить інформацію про здоров'я. Більшість порталів глибинного Web підтримують механізми кластеризації за темами.

Висновки

Однією з причин підвищеного інтересу до проекту Semantic Web є надія на поліпшення пошуку в Web. Дослідження з цієї проблеми ведуться в різних напрямках і дають різноманітні результати у вигляді різних пошукових систем. Такі системи, як Swoogle, дозволяють лише виконувати пошук онтологій за ключовими словами. Але такий сервіс є дуже корисним для розробників семантичних систем і онтологій, хоча він і не розрахований на звичайного користувача. [28]. Джерелами інформації у них служать набори RDF-даних, включаючи дані, що пов'язані в рамках проекту Linked Open Data і мікроформати.

Можна відзначити й інші пошукові системи Semantic Web, багато з яких знаходяться на стадії бета-тестування, тому оцінити їх можливості поки важко. Деякі системи йдуть шляхом „углиблення у Web", інші – більш прискіпливо розвивають алгоритми інтелектуального аналізу та використовують різноманітні джерела інформації про документи, які знаходяться „поза-документом" у Web. Розвиток технологій інформаційного пошуку призвів до інтенсивного використання мета-інформаційно-пошукових систем, багато-агентних інформаційно-пошукових систем, систем, побудованих на реалізації онтологічних, мовних та управлінських угод і їм подібних. Більшість пошукових систем йдуть шляхом розвитку персоналізації пошуку, тобто розпізнання та задоволення потреб користувача.

Традиційні пошукові системи стають все більш точними та об'ємними, однак вони не можуть перевершити інтелект людини. Вони можуть лише порівнювати слова, а не зміст ідеї, яка обговорюється ними. Нові технології пошукових систем 3-го покоління ще знаходяться в стадії формування, але вже нині вони дають по-

зитивні результати. Новий пошук може допомогти зробити пошук більш значущим, суб'єктивним і прив'язаним до задач (task-based), що стоять перед користувачем. Таким чином, розвиток пошукових систем йде в напрямку задоволення потреб окремого користувача, з його перевагами, характером, звичками, поведінкою, рівнем підготовки і знань тощо.

1. *Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze*. An Introduction to Information Retrieval, Online edition (c) 2009 Cambridge UP, Draft of April 1, 2009,Website: http://www.informationretrieval.org
2. *Черний Ю.Ю.* Школа наукової інформації. Інформаційні потреби. Основи інформаційного пошуку, http://www.bogoslov.ru/text/321597.html
3. *Захаров В.П.* Информационно-поисковые системы. Учебно-методическое пособие, Санкт-Петербург, 2005.
4. *Медведь В.Н.* Методы поиска информации, http://northedu.ru/content/view/115/159/
5. *Van Rijsbergen C.J*. Information Retrieval. London: Butterworths, 1979. Available at http://www.dcs.gla.ac.uk/Keith/Preface.html
6. *Шарапов Р.В., Шарапова Е.В., Саратовцева О.А*. Модели информационного поиска.
7. *Некрестьянов И.С.* Тематико-ориентированные методы информационного поиска: Дис. … канд. техн. наук. – Санкт-Петербургский государственный университет. – СПб, 2000. – 88 с.
8. *Дубинский А.Г.* Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. – 2001. – № 4. – С. 77–83.
9. *Когаловский М.Р.* Перспективные технологии информационных систем. – М.: ДМК Пресс; М.: Компания АйТи, 2003. – 288 с.
10. *Holscher C. and Strube G.* Web search behaviour of Internet experts and Newbies. Proceedings of WWW9. 2000. Available at http://www9.org/w9cdrom/81/81.html.
11. *Navarro-Prieto R., Scaife M. & Rogers Y*. Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web, 1999. Available at http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html.
12. *Muramatu J. and Pratt W.* Transparent queries: Investigating Users' Mental Models of Search Engines, Proceedings of SIGIR 2001.
13. *Choo C. W., Detlor B., and Turnbull D*. Information Seeking on the Web – An integrated model of browsing and searching. Proceedings of the Annual Meeting of the American Society for Information Science (ASIS), 1999. Available at http://choo.fis.utoronto.ca/fis/respub/aisis99/
14. *Broder A*. A taxonomy of web search, IBM Research, ACM SIGIR Forum archive. – 2002. – Vol. 36 , Issue 2. – P. 3–10.
15. *Лексична* база англійської мови WordNet, http://wordnet.princeton.edu/perl/webwn
16. *Онлайн* словник http://dictionary.cambridge.org/
17. *Онлайн* словник http://www.merriam-webster.com/
18. *Rodnessey J*. New Search Engines: The Next Generation of Google Competition, 2009, http://webupon.com/search-engines/new-search-engines-the-next-generation-of-google-competition/
19. *Nobles R*. The Future Of Search Engine Optimizing, http://www.searchengineworkshops.com/articles/se-optimization-future.html
20. *Андон Ф.И., Гришанова И.Ю., Резниченко В.А.* Semantic Web как новая модель информационного пространства интернет // Проблеми програмування. – 2008. – № 2–3. – С. 417–430.
21. *Ezzy E*., Search 2.0 vs Traditional Search, 2006, http://www.readwriteweb.com/archives/search_20_vs_tr.php
22. *McLoughlin S*. Searching on the web; the new breed of search engines, 2009, http://relativemusings.blogspot.com/2009/05/searching-on-web-new-breed-of-smarter.html
23. *Wolfram S*. Wolfram Alpha – computational knowledge engine, 2009 http://base-technology.blogspot.com/2009/03/wolfram-alpha-computational-knowledge.html
24. *Сидоров В*. Wolfram Alpha – Computational Knowledge Engine, или Как сложить яблоко с апельсином?, блог, 2009, http://netler.ru/pc/wolfram.htm
25. *Official* Google Blog: Square your search results with Google Squared, http://googleblog.blogspot.com/2009/06/square-your-search-results-with-google.html
26. *Сидоров В*. Google Squared: как успех Wolfram Alpha взбудоражил Google и что из этого вышло?.., блог, 2009, http://netler.ru/pc/google-squared.htm

27. *Soubbotin D.* Summarization, the Answer to Web Search: Interview with Dmitri Soubbotin of SenseBot, Search Engine Journal, 2007, http://www.searchenginejournal.com/summarization-the-answer-to-web-search-interview-with-dmitri-soubbotin-of-sensebot/6094/
28. *Левшин Д.* Web, часть третья // Открытые системы. – 2008. – № 2. http://cio.ru/text/print/302/8165094.html
29. *Рогушина Ю.В., Гришанова І.Ю.* Разработка принципов представления электронных изданий, обеспечивающих корректную индексацию поисковыми системами Интернет // Проблеми програмування. – 2004. – № 4. – С. 39–47.

References

1. *Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze* An Introduction to Information Retrieval, Online edition (c)2009 Cambridge UP, Draft of April 1, 2009, Website: http://www.informationretrieval.org
2. *Cherniy Y.Y.* School of scientific information. Information needs. Basics of information retrieval, http://www.bogoslov.ru/text/321597.html
3. *Zacharov V.P.* Informational retrieval systems, Learning manual, St. Petersburg, 2005
4. *Medvedj V.N.* Methods of information retrieval, http://northedu.ru/content/view/115/159/
5. *Van Rijsbergen C.J.* Information Retrieval. London: Butterworths, 1979. Available at http://www.dcs.gla.ac.uk/Keith/Preface.html
6. *Sharapov P.B.,Sharapova E.V., Saratovceva O.A.*, Models of information retirieval.
7. *Nekrestyanov I.S.* Topic – oriented methods of information retrieval: The Thesis of Ph.D.: 05.13.11 / Saint-Petersburg State University – St.Pt., 2000. – 88 p.
8. *Dubinskyi A.G.* Some questions of the use of the vector model for the document`s presentation in the information retrieval // Control Systems and Computers. – 2001. – N 4. – P. 77–83.
9. *Kogalovskyi M.R.* Prospective technologies of information systems. – M.: DMK Press; Moscow: IT Company, 2003. – 288 p.
10. *Holscher C. and Strube G.* Web search behaviour of Internet experts and Newbies. Proceedings of WWW9. 2000. Available at http://www9.org/w9cdrom/81/81.html.
11. *Navarro-Prieto R., Scaife M. & Rogers Y.* Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web, 1999. Available at http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html.
12. *Muramatu J. and Pratt W.* Transparent queries: Investigating Users' Mental Models of Search Engines, Proceedings of SIGIR 2001.
13. *Choo C. W., Detlor B., and Turnbull D.* Information Seeking on the Web – An integrated model of browsing and searching. Proceedings of the Annual Meeting of the American Society for Information Science (ASIS), 1999. Available at http://choo.fis.utoronto.ca/fis/respub/aisis99/
14. *Broder A.* A taxonomy of web search, IBM Research, ACM SIGIR Forum archive. – 2002. – Vol. 36 , Issue 2. – P. 3–10.
15. *Lexical* base of English language WordNet, http://wordnet.princeton.edu/perl/webwn
16. *Online* vocabulary http://dictionary.cambridge.org/
17. *Online* vocabulary http://www.merriam-webster.com/
18. *Rodnessey J.* New Search Engines: The Next Generation of Google Competition, 2009, http://webupon.com/search-engines/new-search-engines-the-next-generation-of-google-competition/
19. *Nobles R.* The Future Of Search Engine Optimizing, http://www.searchengineworkshops.com/articles/se-optimization-future.html
20. *Andon P.I., Grishanova I.J., Reznichenko V.A.* Semantic Web as a new model of the information space of the Internet // Problems in Programming. – 2008. – N 2–3, P. 417–430.
21. *Ezzy E.* Search 2.0 vs Traditional Search, 2006, http://www.readwriteweb.com/archives/search_20_vs_tr.php
22. *McLoughlin S.* Searching on the web; the new breed of search engines, 2009, http://relativemusings.blogspot.com/2009/05/searching-on-web-new-breed-of-smarter.html
23. *Wolfram S.* Wolfram Alpha – computational knowledge engine, 2009 http://basetechnology.blogspot.com/2009/03/wolfram-alpha-computational-knowledge.html
24. *Sidorov V.* Wolfram Alpha – Computational Knowledge Engine, or How To Add Apple with Orange?, blog, 2009, http://netler.ru/pc/wolfram.htm
25. *Official* Google Blog: Square your search results with Google Squared, http://googleblog.blogspot.com/2009/06/square-your-search-results-with-google.html

26. *Sidorov V.* Google Squared: How the Success of Wolfram Alpha Stirred up Google and What Happened?.., blog, 2009, http://netler.ru/pc/google-squared.htm
27. *Soubbotin D.* Summarization, the Answer to Web Search: Interview with Dmitri Soubbotin of SenseBot, Search Engine Journal, 2007, http://www.searchenginejournal.com/summarization-the-answer-to-web-search-interview-with-dmitri-soubbotin-of-sensebot/6094/
28. *Levshin D.* Web, part 3, "Open systems". – 2008. – N 2. http://cio.ru/text/print/302/8165094.html
29. *Rogushina J.V., Grishanova I.Y.* Development of the Principles of Electronic Publications, Providing the Correct Indexing of Internet Search Engines // Problems in Programming – 2004, N 4. – P. 39–47.

Одержано 08.12.2015

Про автора:

Гришанова Ірина Юріївна,
науковий співробітник,
Кількість наукових публікації в українських виданнях – 15.
http://orcid.org/0000-0003-4999-6294.

Місце роботи автора:

Інститут програмних систем
НАН України,
03181, Київ-187,
Проспект Академіка Глушкова, 40.
E-mail: i26031966@gmail.com