



УКРАЇНА

(19) UA (11) 60952 (13) U
(51) МПК
G06F 17/30 (2006.01)

МІНІСТЕРСТВО ОСВІТИ
І НАУКИ УКРАЇНИ

ДЕРЖАВНИЙ ДЕПАРТАМЕНТ
ІНТЕЛЕКТУАЛЬНОЇ
ВЛАСНОСТІ

ОПИС ДО ПАТЕНТУ НА КОРИСНУ МОДЕЛЬ

видається під
відповідальність
власника
патенту

(54) СПОСІБ ПОШУКУ ІНФОРМАЦІЇ В МАСИВІ ТЕКСТІВ

1

(21) u201104285

(22) 08.04.2011

(24) 25.06.2011

(46) 25.06.2011, Бюл.№ 12, 2011 р.

(72) БАРКОВ АНТОН ЄВГЕНОВИЧ, ГЛАЗУНОВ
ДМИТРО ОЛЕГОВИЧ, ЗЕЛІНСЬКА МАРИНА ОЛЕ-
ГІВНА, ПЕТРУШКЕВИЧ ІРИНА ВІТАЛІЇВНА, ПРА-
ВДИВА ОЛЬГА ВАСИЛІВНА

(73) ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДА-
ЛЬНІСТЮ "ІНФОРМАЦІЙНО-АНАЛІТИЧНИЙ
ЦЕНТР "ЛІГА"

(57) 1. Спосіб пошуку інформації в масиві текстів, згідно з яким за допомогою віддаленого приладу користувача через канали зв'язку пошуковим сервером одержують пошуковий запит із реквізитами та ключовими словами, який проходить аналіз та коригування, далі пошуковий сервер проводить ідентифікацію реквізитів та уточнення запиту, а в кінці обробки проводиться вибір документів зі сховища даних та виведення результатів пошуку за допомогою каналів зв'язку на віддалений прилад користувача, який відрізняється тим, що документи сховища даних складаються з двох частин - тексту документа та картки з його реквізитами, а корегування запиту здійснюють з використанням даних сховища статистичної інформації про картки та тексти документів, далі запит додатково проходить розширення і тільки після цього проводять пошук даних в сховищі даних шляхом обробки змісту запиту по словах, використовуючи попарне співставлення абзаців масиву тестів та пошукового запиту, причому, після пошуку даних в сховищі даних проводять оцінку відповідності реквізитів і ключових слів та оцінку відповідності тексту запиту, далі проводять сортування за відповідністю тексту запиту, наступним етапом є пошук та оцінка місць документів, що містять шукану інформацію, після чого реалізують механізм оцінки, сортування та вибору кращих ідентифікаторів документів визначаючи релевантність результатів пошуку.

2. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що корегування пошукового запиту реалізують через сервер баз даних зі сховищем статистичної інформації про картки та тексти документів.

3. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що віддаленим приладом користувача є: персональний комп'ютер, портатив-

2

ний комп'ютер, планшетний прилад, смартфон та/або будь-який прилад за допомогою якого можливо реалізувати запит інформаційного пошуку.

4. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що каналами зв'язку є Інтернет-мережа, локальна мережа або безпосередньо канали зв'язку пристрою користувача.

5. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що розширення запиту реалізують за допомогою бази даних сховища пов'язаних понять та бази даних зв'язків між картками документів.

6. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що картки та реквізити документа містять інформацію щодо дати, видавника, типу документа, ключових слів тощо, а частини документа пов'язані між собою.

7. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що при пошуку даних в сховищі даних зберігають лише ідентифікатор документа, по якому проводять вибір документів зі сховища даних.

8. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що сховище даних являє собою сукупність пов'язаних між собою логічними й правовими зв'язками документів.

9. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що механізм оцінки включає в себе оцінку ступеня актуальності документів, оцінку за популярністю та цитованістю, оцінку за авторитетністю та оцінку за категорією документів.

10. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що механізм оцінки здійснюється за допомогою сховища відносних оцінок.

11. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що пошук та оцінка місць документів здійснюється за допомогою сховища відносного знаходження в текстах місцезнаходження шуканого.

12. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що сховище даних, в якому здійснюється пошук, являє собою сукупність пов'язаних між собою логічними та правовими зв'язками документів.

13. Спосіб пошуку інформації в масиві текстів, який відрізняється тим, що визначення релевантності результатів пошуку враховує встановлену

(19) UA (11) 60952 (13) U

ієрархію нормативно-правових документів в дер-

жаві.

Корисна модель належить до оброблення цифрових даних за допомогою електричних пристроїв, а саме до способів пошуку інформації в масивах текстів, та може бути застосована, наприклад, для автоматизації пошуку документів нормативно-правової та колоправової тематики в базі даних сховища даних, сформованої певним чином.

В галузі пошуку цифрової інформації в масивах текстів існує суттєва проблема актуальності, точності та швидкості проведення пошукових робіт. Зазвичай це обумовлено недосконалими пошуковими алгоритмами, необхідністю застосування певних, конкретних пристроїв з певним програмним забезпеченням, що дозволяють працювати з певними пошуковими діями, а це звужує коло можливих користувачів та впливає на якість проведення пошукових робіт. Також важливим є надання відібраної інформації, без так званого «шуму» та в логічному порядку, тобто відсіювати неактуальну та застарілу інформацію та сортувати її в зручному для розуміння користувачу порядку та обсязі.

Існує відомий спосіб введення та пошуку інформації про об'єкт у віддаленій базі даних [патент України на винахід № 92117 «Спосіб введення та пошуку інформації про об'єкт у віддаленій базі даних», дата подання 04. 06. 2007, опубліковано 27. 09. 2010, Бюл. № 18, 2010 р.], що включає передачу за допомогою мережі глобальної системи мобільного зв'язку через устаткування оператора стільникового зв'язку та за допомогою мережі Інтернет на сервер від мобільного термінала зв'язку повідомлення з ідентифікатором об'єкта, прийом сервером повідомлення з ідентифікатором об'єкта, пошук ідентифікатора об'єкта у базі даних сервера, а при його знаходженні передачу від сервера повідомлення мобільному терміналу зв'язку про знаходження ідентифікатора об'єкта і, згідно з винаходом, на мобільному терміналі зв'язку та сервері встановлюють програмне забезпечення, що дозволяє як ідентифікатор об'єкта використовувати слово, а як повідомлення з ідентифікатором об'єкта, що передане від мобільного термінала зв'язку, використовують щонайменше одне слово, що характеризує найменування/діяльність юридичної/фізичної особи, при пошуку ідентифікатора об'єкта у базі даних сервера сервер зіставляє слово з доменними іменами, що зберігаються у його базі даних, і при виявленні доменних імен, принаймні частина яких збігається зі словом, передає на мобільний термінал зв'язку список з доменними іменами, при виборі користувачем на мобільному терміналі зв'язку одного доменного імені зі списку користувач за допомогою мобільного термінала зв'язку формує повідомлення з вибраним доменним іменем і передає його на сервер, який проводить додатковий пошук цього доменного імені та відповідного йому телефонного номера абонента, при знаходженні відповідного телефонного номера абонента сервер передає цей номер

на мобільний термінал зв'язку, при підтвердженні вибору користувачем на мобільному терміналі зв'язку цього доменного імені/телефонного номера абонента програмне забезпечення за допомогою мобільного термінала зв'язку ініціює зв'язок через глобальну систему мобільного зв'язку та устаткування оператора стільникового зв'язку з телефонним номером абонента, здійснюючи через них голосове з'єднання з телефонним номером абонента.

Недоліками відомого способу є обов'язкове застосування пристроїв мобільного зв'язку та мобільних терміналів, що обмежує обсяг інформації та швидкість проведення пошукових робіт.

Найбільш близьким до запропонованого є спосіб пошуку інформаційних об'єктів [описаний в патенті України на винахід № 90764, дата подання 13. 05. 2008, опубліковано 25. 05. 2010, бюл. № 10, 2010 р. «СПОСІБ ПОШУКУ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ ТА СИСТЕМА ДЛЯ ЙОГО ЗДІЙСНЕННЯ»], згідно з яким приймають запит, здійснюють збір інформації, пошук по інформаційному сховищу на предмет об'єктів, що відповідають запиту користувача на пошук і, згідно з винаходом, включає визначення елементів уточнення на основі щонайменше одного з рейтингу елементів уточнення та інформації про елементи уточнення, після чого вибирають елементи уточнення, включають вибрані елементи в запит, остаточно редагують та підтверджують запит, визначають еквіваленти дескрипторів та елементів уточнення, отримують від користувача команду на підключення інтеграції запиту, обробляють пошуковою системою запит шляхом вибору документів або текстів із сховища даних та здійснюють виведення результатів.

Недоліками найбільш близького до запропонованого способу є недостатня точність проведення пошукових робіт та відносно довгий час обробки інформації.

Задачею корисної моделі є створення способу пошуку інформації в масивах текстів в базі даних сховища даних, який би розширював функціональні можливості та підвищував техніко-експлуатаційні характеристики пошукових робіт, підвищував швидкість, актуальність та чистоту пошуку. Також важливим питанням, що вирішує запропонований спосіб є визначення релевантності для результатів пошуку.

Поставлена задача досягається завдяки запропонованому способу, згідно з яким: за допомогою віддаленого приладу користувача через канали зв'язку пошуковим сервером одержують пошуковий запит із реквізитами та ключовими словами, який проходить аналіз та коригування, далі пошуковий сервер проводить ідентифікацію реквізитів та уточнення запиту і, згідно з корисною моделлю, документи сховища даних складаються з двох частин - тексту документа та картки з його реквізитами, а корегування запиту здійснюють з

використанням даних сховища статистичної інформації про картки та тексти документів, далі запит додатково проходить розширення і тільки після цього проводять пошук даних в сховищі даних шляхом обробки змісту запиту по словах, використовуючи попарне співставлення абзаців масиву тестів та пошукового запиту, причому, після пошуку даних в сховищі даних проводять оцінку відповідності реквізитів і ключових слів та оцінку відповідності тексту запиту, далі проводять сортування за відповідністю тексту запиту, наступним етапом є пошук та оцінка місць документів, що містять шукану інформацію, після чого реалізують механізм оцінки, сортування та вибору кращих ідентифікаторів документів, далі проводиться вибір документів зі сховища даних та виведення результатів пошуку за допомогою каналів зв'язку на віддалений прилад користувача визначаючи та враховуючи релевантність.

Корегування пошукового запиту реалізують через сервер баз даних зі сховищем статистичної інформації про картки та тексти документів. Віддаленим приладом користувача є: персональний комп'ютер, портативний комп'ютер, планшетний прилад, смартфон та/або будь-який прилад, за допомогою якого можливо реалізувати запит інформаційного пошуку. Каналами зв'язку є Інтернет-мережа, локальна мережа або безпосередньо канали зв'язку пристрою користувача. Розширення запиту реалізують за допомогою бази даних сховища пов'язаних понять та бази даних зв'язків між картками документів. Картки та реквізити документа містять інформацію щодо дати, видавника, типу документа, ключових слів тощо, а частини документа пов'язані між собою. При пошуку даних в сховищі даних зберігають лише ідентифікатор документа, по якому проводять вибір документів зі сховища даних. Сховище даних являє собою сукупність пов'язаних між собою логічними й правовими зв'язками документів. Механізм оцінки включає в себе оцінку ступеня актуальності документів, оцінку за популярністю та цитованістю, оцінку за авторитетністю та оцінку за категорією документів. Механізм оцінки здійснюється за допомогою сховища відносних оцінок, а пошук та оцінка місць документів здійснюється за допомогою сховища відносного знаходження в текстах місцезнаходження шуканого. Сховище даних, в якому здійснюється пошук, являє собою сукупність пов'язаних між собою логічними та правовими зв'язками документів. Визначення релевантності результатів пошуку враховує встановлену ієрархію нормативно-правових документів в державі.

Корисна модель пояснюється кресленням, на якому зображено:

- загальна блок-схема способу пошуку інформації в масиві текстів.

Спосіб здійснюють наступним чином:

за допомогою віддаленого приладу користувача 1 через канали зв'язку 2 пошуковий сервер 3 одержує пошуковий запит 4 із реквізитами та ключовими словами, який проходить аналіз 5 та коригування 6, далі пошуковий сервер 3 проводить ідентифікацію 7 реквізитів та уточнення запиту 8. Згідно з корисною моделлю, документи сховища

даних 9 складаються з двох частин - тексту документа та картки з його реквізитами, а корегування 6 запиту здійснюють з використанням даних сховища статистичної інформації 10 про картки та тексти документів, далі запит додатково проходить розширення 11 і тільки після цього проводять пошук даних 12 в сховищі даних 9 шляхом обробки змісту запиту по словах, використовуючи попарне співставлення абзаців масиву тестів та пошукового запиту, причому, після пошуку даних 12 в сховищі даних 9 проводять оцінку відповідності реквізитів і ключових слів 13 та оцінку відповідності тексту запиту 14, далі проводять сортування 15 за відповідністю тексту запиту, наступним етапом є пошук та оцінка місць документів 16, що містять шукану інформацію, після чого реалізують механізм оцінки 17, який включає в себе оцінку ступеня актуальності документів 21, оцінку за популярністю та цитованістю 22, оцінку за авторитетністю 23 та оцінку за категорією документів 24. Далі проводять сортування 18 та вибір 19 кращих ідентифікаторів документів, після чого проводиться вибір документів зі сховища даних 9 та виведення результатів 20 пошуку за допомогою каналів зв'язку 2 на віддалений прилад користувача 1. Віддаленим приладом користувача 1 є: персональний комп'ютер, портативний комп'ютер, планшетний прилад, смартфон та/або будь-який прилад, за допомогою якого можливо реалізувати запит інформаційного пошуку. Каналами зв'язку 2 є інтернет-мережа, локальна мережа або безпосередньо канали зв'язку пристрою користувача. Розширення запиту 4 реалізують за допомогою сховища пов'язаних понять 25 та бази даних зв'язків між картками документів 26. Картки та реквізити документа містять інформацію щодо дати, видавника, типу документа, ключових слів тощо, а частини документа пов'язані між собою. При пошуку даних в сховищі даних зберігають лише ідентифікатор документа, по якому проводять вибір документів зі сховища даних. Сховище даних 9 являє собою сукупність пов'язаних між собою логічними й правовими зв'язками документів. Механізм оцінки 17 здійснюється за допомогою сховища відносних оцінок 27, а пошук та оцінка місць документів 16 здійснюється за допомогою сховища відносного знаходження в текстах місцезнаходження шуканого 28.

Визначення релевантності результатів пошуку при запитах користувача відіграє важливу роль при аналізі користувачем отриманої, в результаті пошуку, інформації. Важливо, щоб користувач в першу чергу побачив найактуальнішу та вагомішу на цей час інформацію з усього результату пошуку. При визначенні релевантності в даному способу беруться до уваги специфічні оцінки документів, що визначають значимість для певного запиту, а це тип документа, видавник документа, дата документа, актуальність, популярність та цитованість документа, оцінка авторитетності органа, що видав документ, та категорії нормативно-правових документів. Всі ці дії та наявність необхідної інформації в сховищах даних дозволяють отримувати результати пошуку в актуальному порядку, що значно полегшує роботу користувача з отриманими результатами пошуку.

Таким чином завдяки розробленому та запропонованому способу пошуку інформації в масивах текстів в сховищах даних досягається розширення функціональних можливостей пошуку, підвищується

техніко-експлуатаційні характеристики, зменшується час на обробку інформації, покращується актуальність та чистота пошуку.

