УКРАЇНА **(19) UA** **(11) 135223** **(13) U**

(51) МПК
G06F 17/21 (2006.01)
G06F 17/27 (2006.01)
G06F 17/28 (2006.01)

МІНІСТЕРСТВО ЕКОНОМІЧНОГО РОЗВИТКУ І ТОРГІВЛІ УКРАЇНИ

(12) ОПИС ДО ПАТЕНТУ НА КОРИСНУ МОДЕЛЬ

(21) Номер заявки: **u 2019 00016**
(22) Дата подання заявки: **02.01.2019**
(24) Дата, з якої є чинними права на корисну модель: **25.06.2019**
(46) Публікація відомостей про видачу патенту: **25.06.2019, Бюл.№ 12**

(72) Винахідник(и):
Бісікало Олег Володимирович (UA),
Лісовенко Анна Ігорівна (UA),
Яхимович Олександр Вікторович (UA),
Шолота Владислава Владиславівна (UA)

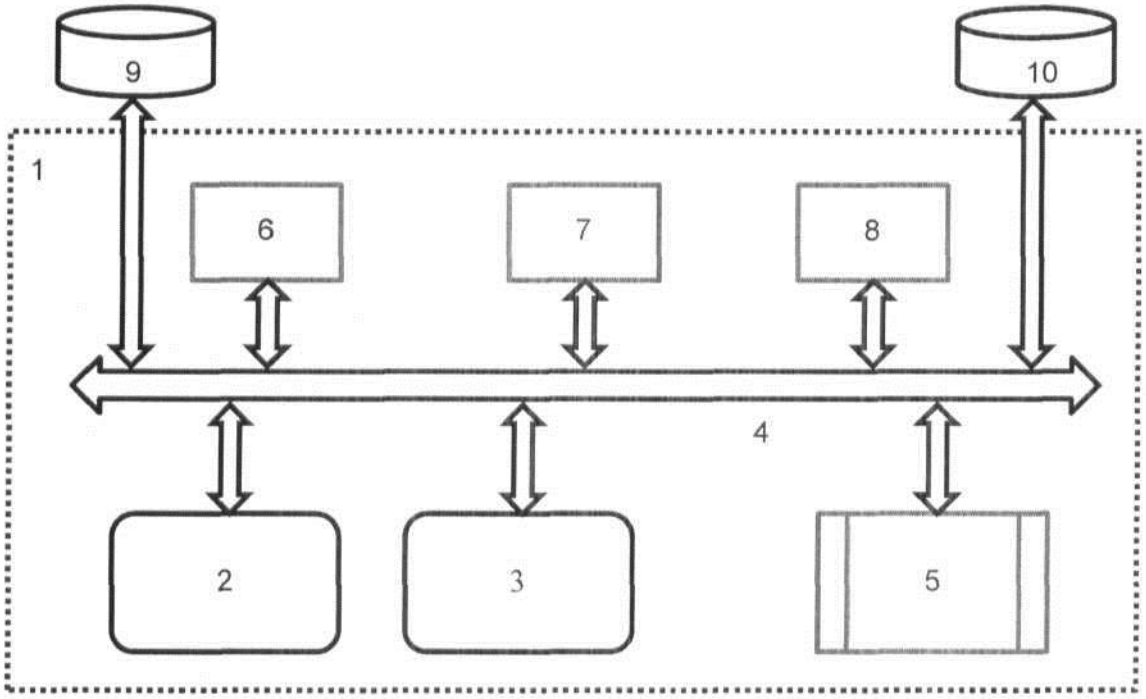
(73) Власник(и):
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ,
Хмельницьке шосе, 95, м. Вінниця, 21021 (UA)

(54) СПОСІБ АВТОМАТИЧНОГО ПОШУКУ КЛЮЧОВИХ СЛІВ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЇ DKPro Core

(57) Реферат:

Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core включає знаходження словосполучень кандидатів зі застосуванням аналізатора DKPro, розбір словосполучень кандидатів в набір слів, побудову семантичного графа між словосполученнями кандидатів, відбір найпопулярніших кандидатів як ключових слів на основі кількості зв'язків між словосполученнями кандидатів, в яких ключові слова отримані шляхом обробки семантики, причому словосполучення кандидатів у ключові слова виокремлюють з текстового документу, що зберігається в колекції текстів на зовнішньому запам'ятовуючому пристрої, за допомогою аналізатора DKPro текст перетворюють у набір речень і зв'язків між членами цих речень, а кожне речення розбивають на словосполучення кандидатів, для яких, використовуючи персональний комп'ютер, визначають головне і залежне слово, а також тип зв'язку між ними, за допомогою програмного модуля виключають словосполучення кандидатів, зв'язки яких внесені у попередньо заданий перелік не інформативних для семантичного аналізу типів зв'язків та будують семантичний граф між словосполученнями кандидатів, за допомогою програмного модуля, використовуючи апаратні складові персонального комп'ютера, займенники в словосполученнях кандидатів замінюють на відповідні до них іменники, а словосполучення кандидатів розбивають на окремі слова, для кожного з яких за допомогою програмного модуля знаходять частину мови І лему, а також визначають кількість семантичних зв'язків для окремого слова.

UA 135223 U


9
10
1
6
7
8
4
2
3
5

Корисна модель належить до галузі комп'ютерної лінгвістики.

Відомий спосіб визначення ключових слів [заявка на патент США№ 20140289260, м. кл. G06F17/30, опубл. 25.09.2014].

Спосіб полягає у знаходженні короткого змісту тексту та ідентифікації ключових слів, пов'язаних із даним текстом. Визначення списку ключових слів відбувається на основі порівняння частоти появи ключового слова в короткому змісті тексту та частоти появи слова у тексті в цілому. Вибір ключових слів відбувається на основі певного граничного значення. Якщо частота зустрічі слова у тексті перевищує визначене граничне значення, дане слово є ключовим.

Недоліком описаного способу є неможливість визначення користувачем потрібної йому кількості ключових слів.

Відомий спосіб вилучення ключових слів з тексту [патент ЄС № 0364179, м. кл. G06F17/30, G06F17/27, опубл. 18.05.1990].

Спосіб полягає у автоматичному вилученні ключових слів із файлу. Це може бути корисно при роботі з файлами у великих файлових системах. Для визначення ключового слова, обчислюється відношення частоти появи кожного слова у файлі до частоти виникнення інших слів у ньому ж. Якщо розрахована частота перевищує деяке задане граничне значення частоти появи цього ж слова у довідковій області, що відповідає файлу, то слово визначається як ключове для даного файлу.

Недоліком описаного способу є відсутність механізму, що визначає змістовне навантаження слова, тобто за даним способом немає можливості відсіяти службові частини мови та інші високочастотні слова.

Відомий спосіб вилучення ключових слів з тексту, описаний в заявці на винахід Росії № 2004114529, м. кл. G06F17/21, опубл. 27.10.2005.

Спосіб полягає у автоматичному пошуку за допомогою ЕОМ ключових слів і словосполучень в сегментованому тексті на мові з ідеографічною системою письма, що включає в себе частотний і дистрибутивний методи визначення значущих одиниць тексту, на основі яких проводять пошук ключових слів і словосполучень. Як основні елементи аналізу використовують знаки системи писемності ідеографічних мов - ієрогліфи, при цьому кожну одиницю тексту (абзац, речення, синтагму) індексують. На основі отриманого індексу знаходять слова, що являють собою послідовності знаків у межах однієї синтагми, що зустрічаються в тексті не менше двох разів. Отримані слова порівнюють за абсолютною частотою для даного тексту, при цьому слова із частотою, що перевищує певний поріг (наприклад, середнє значенні абсолютної частоти всіх слів отриманого списку), є ключовими словами.

Недоліком вказаного способу є те, що не враховуються синтаксичні зв'язки між словами в тексті, які мають інформативне значення. Це приводить до зменшення швидкодії та релевантності.

Найближчим аналогом є спосіб знаходження ключових слів для реклами за допомогою семантики онлайн-енциклопедії [патент США № 8768960 B2, м. кл. G06F17/30, опубл. 1.07.2014].

Спосіб включає: знаходження словосполучень кандидатів з застосуванням аналізатора для перетворення веб-сторінки в дерево DOM, лексичний розбір вмісту дерева DOM в набір слів, і використання індексу словосполучень онлайн-енциклопедії для визначення, чи є слово з набору слів пов'язаним з одним або декількома записами в онлайн-енциклопедії; генерують індекси словосполучень онлайн-енциклопедії, що складається з одного або декількох наступних елементів: статей, заголовків перенаправлених сторінок, сторінок тлумачень та прив'язок (якорів) в тексті; визначають категорії, що містять записи в онлайн-енциклопедії, шляхом збору гіпер-категорії записів онлайн-енциклопедії; установлюють зв'язки між словосполученнями кандидатами і записами з онлайн-енциклопедії: установлення словосполученню кандидату одного запису з онлайн-енциклопедії; визначають поняття, що відповідають словосполученням: встановлюють однозначну відповідність між множиною записів онлайн-енциклопедії, на основі, принаймні частково, одного або обох подібностей тексту між веб-сторінкою і множиною записів онлайн-енциклопедії або категоріальну відповідність між тегом/тегами категорії словосполучень кандидатів та категорій записів онлайн-енциклопедії; побудову семантичного дводольного графа між словосполученнями кандидатів та категоріями, що містять записи в онлайн-енциклопедії; відбір одного або декількох словосполучень кандидатів як ключових слів з використання алгоритму пошуку HITS (Hyperlink-Induced Topic Search) в семантичному дводольному графі, щоб вибрати N найпопулярніших словосполучень кандидатів і категорій як ключових слів і тем для веб-сторінки, на основі, принаймні частково, кількості зв'язків між словосполученнями кандидатами і категоріями словосполучень кандидатів, категорії,

включають записи з онлайн-енциклопедії, в яких ключові слова для цільової реклами отримані шляхом обробки семантики, на основі таксономії онлайн-енциклопедії; доповнюють ключові слова застосувавши статистичний екстрактор словосполучень для веб-сторінки для здійснення контрольованого пошуку за ключовими словами.

Недоліками даного способу є низька швидкодія, обумовлена необхідністю аналізу всіх записів онлайн-енциклопедії, щоб визначити ключові слова для одного тексту, а також проводити такий аналіз регулярно, бо з'являється велика кількість нових статей, що призводить до постійних затрат ресурсів на аналіз всіх текстів корпусу.

В основу корисної моделі поставлено задачу створення способу автоматичного пошуку ключових слів з використанням технології DKPro Core, в якому за рахунок введення нових операцій та їх послідовності досягається можливість визначати ключові слова для одного тексту без аналізу всіх текстів корпусу, що призводить до збільшення швидкодії.

Згідно з корисною моделлю в способі автоматичного пошуку ключових слів з використанням технології DKPro Core, що включає знаходження словосполучень кандидатів зі застосуванням аналізатора DKPro, розбір словосполучень кандидатів в набір слів, побудову семантичного графа між словосполученнями кандидатів, відбір найпопулярніших кандидатів як ключових слів на основі кількості зв'язків між словосполученнями кандидатів, в яких ключові слова отримані шляхом обробки семантики, словосполучення кандидатів у ключові слова виокремлюють з текстового документу, що зберігається в колекції текстів на зовнішньому запам'ятовуючому пристрої, за допомогою аналізатора DKPro текст перетворюють у набір речень і зв'язків між членами цих речень, а кожне речення розбивають на словосполучення кандидатів, для яких, використовуючи персональний комп'ютер (ПК), визначають головне і залежне слово, а також тип зв'язку між ними, за допомогою програмного модуля виключають словосполучення кандидатів, зв'язки яких внесені у попередньо заданий перелік не інформативних для семантичного аналізу типів зв'язків та будують семантичний граф між словосполученнями кандидатів, за допомогою програмного модуля, використовуючи апаратні складові ПК, займенники в словосполученнях кандидатів замінюють на відповідні до них іменники, а словосполучення кандидатів розбивають на окремі слова, для кожного з яких за допомогою програмного модуля знаходять частину мови і лему, а також визначають кількість семантичних зв'язків для окремого слова.

На кресленні представлено схему пристрою, за допомогою якого здійснюється спосіб автоматичного пошуку ключових слів з використанням технології DKPRO CORE.

Пристрій складається з ПК 1 з апаратними складовими - центральний процесор 2, запам'ятовуючі пристрої (ОЗП та ПЗП - 3), загальна шина даних (канал зчитування та передачі даних - 4), а також програмні складові, необхідні для реалізації способу, зокрема аналізатор DKPro 5, здатний виокремити з тексту кожне речення і розбити його на словосполучення кандидатів, для яких визначити головне і залежне слово та зв'язок між ними, модуль виключення словосполучень кандидатів з не інформативними типами зв'язків 6, модуль заміни займенників на іменники в словосполученнях кандидатів 7, модуль визначення частини мови, леми і кількості семантичних зв'язків для кожного окремого слова зі словосполучень кандидатів 8. До пристрою приєднуються два зовнішніх запам'ятовуючих пристрої, де зберігаються колекція текстів 9 та ключові слова 10.

Спосіб здійснюється наступним чином.

На вхід каналу зчитування та передачі даних 4 надходить текст і перелік не інформативних для семантичного аналізу типів зв'язків, що розміщені на зовнішньому запам'ятовуючому пристрої, де зберігається колекція текстів 9.

За допомогою аналізатора DKPro 5 текст з колекції текстів 9 перетворюють в набір речень і зв'язків між членами цих речень, а кожне речення розбивають на словосполучення кандидатів у ключові слова, для кожного словосполучення, використовуючи ПК 1, визначають головне і залежне слово, а також тип зв'язку між ними, нарівні центрального процесора 2 будують семантичний граф між словосполученнями кандидатів, який зберігається на запам'ятовуючих пристроях 3.

У модулі виключення словосполучень кандидатів з неінформативними типами зв'язків 6 на рівні центрального процесора 2 перевіряється тип зв'язку кожного словосполучення. Якщо тип зв'язку належить до попередньо заданого переліку не інформативних для семантичного аналізу типів зв'язків, таке словосполучення видаляється, а ті словосполучення кандидатів у ключові слова, що залишилися, зберігаються на запам'ятовуючих пристроях 3.

У модулі заміни займенників на іменники 7 на рівні центрального процесора 2 у словосполученнях кандидатів знаходять займенники і перелік іменників тексту, які потенційно відповідають цим займенникам, за допомогою аналізатора DKPro 5 замінюють займенники в

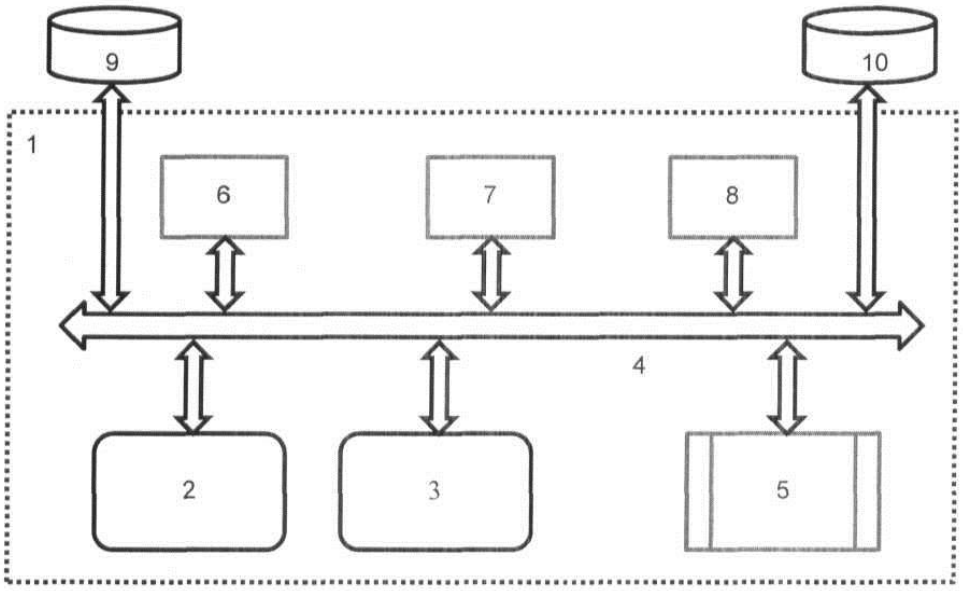
словосполученнях кандидатів, що зберігаються на запам'ятовуючих пристроях 3, на відповідні до них іменники.

Далі, у модулі визначення частин мови, лем і кількості семантичних зв'язків 8 словосполучення кандидатів на рівні центрального процесора 2 розбивають на окремі слова, для кожного з яких за допомогою аналізатора DKPro 5 знаходять частину мови і лему, а також визначають кількість семантичних зв'язків для кожного окремого слова, отримана інформація зберігається на запам'ятовуючих пристроях 3.

Отримані ключові слова на завершальному етапі сортуються за кількістю семантичних зв'язків на рівні центрального процесора 2, по каналу зчитування та передачі даних 4 передаються на зовнішній запам'ятовуючий пристрій 10, де і зберігаються.

ФОРМУЛА КОРИСНОЇ МОДЕЛІ

Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core, що включає знаходження словосполучень кандидатів зі застосуванням аналізатора DKPro, розбір словосполучень кандидатів в набір слів, побудову семантичного графа між словосполученнями кандидатів, відбір найпопулярніших кандидатів як ключових слів на основі кількості зв'язків між словосполученнями кандидатів, в яких ключові слова отримані шляхом обробки семантики, який **відрізняється** тим, що словосполучення кандидатів у ключові слова виокремлюють з текстового документу, що зберігається в колекції текстів на зовнішньому запам'ятовуючому пристрої, за допомогою аналізатора DKPro текст перетворюють у набір речень і зв'язків між членами цих речень, а кожне речення розбивають на словосполучення кандидатів, для яких, використовуючи персональний комп'ютер, визначають головне і залежне слово, а також тип зв'язку між ними, за допомогою програмного модуля виключають словосполучення кандидатів, зв'язки яких внесені у попередньо заданий перелік не інформативних для семантичного аналізу типів зв'язків та будують семантичний граф між словосполученнями кандидатів, за допомогою програмного модуля, використовуючи апаратні складові персонального комп'ютера, займенники в словосполученнях кандидатів замінюють на відповідні до них іменники, а словосполучення кандидатів розбивають на окремі слова, для кожного з яких за допомогою програмного модуля знаходять частину мови І лему, а також визначають кількість семантичних зв'язків для окремого слова.



Комп'ютерна верстка Г. Паяльніков

Міністерство економічного розвитку і торгівлі України, вул. М. Грушевського, 12/2, м. Київ, 01008, Україна

ДП "Український інститут промислової власності", вул. Глазунова, 1, м. Київ – 42, 01601