

Облікова картка ДіР



I. Загальні відомості

Державний обліковий номер: 0219U003681

Державний реєстраційний номер: 0117U004726

Особливі позначки: відкрита

Дата реєстрації: 08-02-2019

II. Етап виконання ДіР

Номер етапу: 1

Назва етапу: Розробка моделей та методів ідентифікації процесів екстракції знань у слабко структурованих масивах інформації

Початок етапу: 01.2017

Закінчення етапу: 12.2018

Вид звітнього документа: Остаточний звіт

III. Відомості про виконавця ДіР

Повне найменування юридичної особи (або ПІБ фізичної особи): Національний технічний університет "Харківський політехнічний інститут"

Код за ЄДРПОУ: 02071180

Місцезнаходження:

Форма власності:

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: Не застосовується

Розмір організації:

Телефон: +38(057)707-63-60

IV. Відомості про співвиконавців ДіР

V. Відомості про замовника ДіР

Повне найменування юридичної особи (або ПІБ фізичної особи): Національний технічний університет "Харківський політехнічний інститут"

Код за ЄДРПОУ: 02071180

Місцезнаходження: вул. Кирпичова, 2, м. Харків, Харківський р-н., Харківська обл., 61002, Україна

Форма власності:

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: Не застосовується

Розмір організації:

Телефон: 0577001564

VI. Джерела, напрями та обсяги фінансування ДіР

Підстава для проведення ДіР: 43 - власна ініціатива (якщо робота виконується з власної ініціативи за кошти виконавця НДР або безкоштовно)

Напрямок фінансування: 2.2 - прикладні дослідження і розробки

Джерела фінансування

7706 - безплатно (договір про науково-технічне співробітництво, тощо)

VII. Відомості про ДіР

Назва роботи українською:

Розробка моделей та методів ідентифікації процесів екстракції знань у слабо структурованих масивах інформації

Назва роботи англійською:

Development of models and methods for identifying knowledge extraction processes in weakly structured information arrays

Реферат українською:

Об'єкт дослідження - процеси екстракції знань у слабо структурованих текстових масивах інформації. Мета роботи - розробка моделей, методів та інформаційних технологій інтелектуальної обробки даних та предметних знань для автоматизованої ідентифікації процесів екстракції знань у слабо структурованих масивах інформації. Методи дослідження базуються на комплексному використанні методів теорії інтелекту, апарату алгебри відношень та алгебри операцій над відношеннями, алгебри предикатних операцій та алгебри підстановних операцій, методу компараторної ідентифікації, теоретичних засад та практичних методів комп'ютерної лексикографії, квантитативної та корпусної лінгвістики. Розроблено методи та створено інформаційні технології автоматичного видобування з текстів предметної області лексикографічних одиниць та іменованих сутностей. Створено робочий корпус для вилучення іменованих сутностей української мови. Розроблено спеціальне інформаційно-лексикографічне забезпечення вирішення задачі вилучення іменованих сутностей української мови. Для проведення експериментальних досліджень створено прототипи програм для автоматичного вилучення лексикографічних одиниць та іменованих сутностей. Оцінку ефективності здійснено окремо для двох основних задач, які вирішуються у дослідженні: задачі видобування термінів предметної області з текстів та задачі видобування іменованих сутностей. Для першої задачі Recall = 0,89, Precision = 0,94, для другої - Recall = 0,67, Precision = 0,89. Порівняння отриманих результатів з результатами подібних систем показали ефективність роботи інформаційних систем щодо вирішення представлених двох задач, а також підтверджують відмінні переваги отриманих результатів запропонованих моделей та інформаційних технологій у проведеному аналізі в порівнянні з вітчизняними й зарубіжними аналогами. Отримані результати можуть бути використані у вигляді математичного, алгоритмічного, інформаційного, програмного та іншого забезпечення системи автоматизованого створення лексикографічних ресурсів різного призначення, а також системи автоматизованого видобування іменованих сутностей. Одержані у дослідженні результати впроваджено у вигляді пакетів прикладних програм для створення галузевих інформаційних ресурсів, у тому числі лексикографічних, здійснення тематичного, патентного та інформаційного пошуку.

Реферат англійською:

The object of the study is the processes of extraction of knowledge in weakly structured text arrays of information. The purpose of the work is the development of models, methods and information technologies for the intellectual processing of data and subject knowledge for the automated identification of extraction processes of knowledge in weakly structured arrays of information. The research methods are based on the integrated use of the methods of the theory of intelligence, the apparatus of the algebra of relations and the algebra of operations on relations, algebra of predicate operations and the algebra of subtractive operations, the method of comparative identification, theoretical foundations and practical methods of computer lexicography, quantitative and corpus linguistics. Methods are developed and information technologies of automatic extraction from the texts of the subject area of lexicographic units and named entities are created. A workshop was created to remove the Ukrainian-language entities. A special information-lexicographic support for solving the problem of extracting the named entities of the Ukrainian language has been developed. For experimental research, prototypes of programs

for the automatic extraction of lexicographic units and named entities are created. The evaluation of efficiency is carried out separately for two main tasks that are solved in the study: tasks of extracting the terms of the subject domain from the texts and the task of extracting named entities. For the first problem, Recall = 0.89, Precision = 0.94, for the second - Recall = 0.67, Precision = 0.89. Comparison of the results with the results of similar systems showed the effectiveness of the work of information systems in solving the two problems presented, and also confirm the excellent advantages of the results of the proposed models and information technologies in the analysis compared with domestic and foreign analogues. The obtained results can be used in the form of mathematical, algorithmic, informational, software and other support of the system of automated creation of lexicographic resources of various purposes, as well as systems of automated mining of named entities. The results obtained in the study are implemented as packages of applications for the creation of industry information resources, including lexicographic, thematic, patent and information retrieval.

Індекс УДК: 004.89:004.93, 004.89:004.93

Коди тематичних рубрик: 28.23.24

Керівники роботи

Власне Прізвище Ім'я По-батькові: Шаронова Наталія Валеріївна

Науковий ступінь:

Вчене звання:

Ідентифікатор ORCID ID:

Додаткова інформація:

VIII. Наукова (науково-технічна) продукція (НТП)

Назва НТП українською: Моделі, методи та інформаційні технології інтелектуальної обробки даних та предметних знань на основі автоматичної ідентифікації іменованих сутностей української мови, а також автоматичної побудови парадигматичних та синтагматичних відношень лексикографічних одиниць української мови, автоматичної ідентифікації структурно-граматичних особливостей термінів предметної області.

Назва НТП англійською: Models, methods and information technologies of intelligent data and subject knowledge processing on the basis of automatic named entities of the Ukrainian language identification, as well as automatic construction of paradigmatic and syntagmatic relations of lexicographic units of the Ukrainian language, automatic identification of structural and grammatical features of the terms of the subject area.

НТП, яку передбачалося створити:

Причини, через які НТП не було створено:

Отримані результати:

Галузь застосування: 72.10

Реєстраційний номер картки технології:

Опис НТП: Розроблено інформаційну систему автоматизованого формування лексикографічних ресурсів, а саме термінологічного словника предметної області та словника синонімів термінів предметної області, система також виконує задачу автоматизованого видобування термінів предметної області з текстів. Розроблено інформаційну систему автоматизованого видобування іменованих сутностей з текстів. Інформаційні системи можуть використовуватися для створення галузевих інформаційних ресурсів, у тому числі лексикографічних, здійснення тематичного, патентного та інформаційного пошуку.

Соціально-економічна спрямованість НТП:

Вплив НТП на довкілля:

Впровадження НТП: Не впроваджено

Практична реалізація НТП

Початок етапу: 2019 р.

Закінчення етапу:

Споживачі продукції: ІТ-компанії, що розробляють програмне забезпечення для автоматизованої обробки природної мови

Перспективні ринки: ринки України

Характер співробітництва з інвестором

Потрібний обсяг інвестицій, тис. грн.:

Права, що надаються інвестору після завершення роботи:

Наявність бізнес-плану:

Техніко-економічне обґрунтування:

Потенціальний обсяг продажу, тис. грн.:

Очікуваний термін окупності (років):

Додаткова інформація:

ІХ. Бібліографічний опис

Х. Заключні відомості

Керівник юридичної особи

Лісачук Георгій Вікторович

Перелік осіб-виконавців

Бабкова Надія Вікторівна

Борисова Наталя Володимирівна

Гулієва Діна Олександрівна

Каніщева Ольга Валеріївна

Кочуєва Зоя Анатоліївна

Купріянов Євген Валерійович

Оробінська Олена Олександрівна

Петрасова Світлана Валентинівна

Хайрова Ніна Феліксівна

Шабанова-Кушнарєнко Любов Володимирівна

Шаронова Наталія Валеріївна

Відповідальний за підготовку облікових документів

Телефон

Реєстратор

Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності

