РЕФЕРАТ

Магістерська дисертація: 99 с., 19 рис., 22 табл., 1 додаток, 63 джерела.

Актуальність. Потреба в автоматичній обробці текстових документів зараз є надзвичайно високою, і постійно зростає. Це обумовлено щоденним збільшенням текстової інформації на просторах всесвітньої мережі інтернет. За даними на березень 2016 року в Інтернеті знаходиться близько 4,66 млрд сторінок, при чому ця цифра включає лише сторінки, які індексовані в найбільш розповсюджених пошукових системах. Тож, без комп'ютерної обробки виконати аналіз такого об'єму інформації за прийнятний час не можливо.

Одною із задач інтелектуального аналізу текстів є їх класифікація на задані категорії, яка потребує вирішення в різних сферах людської діяльності. Так, для забезпечення інформаційної та суспільної безпеки, важливе значення має аналіз даних соціальних мереж, блогів тощо, з метою виявлення даних пов'язаних з тероризмом, наркоторгівлею і т.д. Також в комерційній та суспільній діяльності часто постає потреба обробки відгуків та коментарів, з метою виявлення їх емоційного забарвлення (негативного або позитивного), їх розподіл на подальше опрацювання між різними підрозділами і т.д. В першому та в другому прикладах постає задача класифікації текстової інформації між категоріями в умовах обмеженості за часом та ресурсами обчислювальних пристроїв. Тому задача автоматичної та якісної класифікації даних за прийнятний час, без попереднього структурування інформації, оскільки структурування потребує додаткових ресурсів, часу та не завжди може пройти без втрати важливої інформації, є задачею, яка варта уваги та досліджень.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалась на кафедрі автоматизованих систем обробки інформації та управління Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського» в рамках теми «Методи та технології високопродуктивних обчислень та обробки надвеликих масивів даних». Державний реєстраційний номер 0117U000924.

Метою дослідження є покращення якості моніторингу медіа активності шляхом розробки алгоритму автоматичного аналізу текстової інформації, що дозволяє підвищити точність та повноту класифікації.

Для досягнення поставленої мети необхідно виконати наступні завдання:

– проаналізувати алгоритми та методи машинного навчання для вирішення задачі автоматичної класифікації текстів;

– обрати модель представлення текстової інформації в класифікаторі;

– розробити алгоритм попередньої обробки текстів відповідно до обраної моделі представлення текстової інформації;

– розробити модифікований метод класифікації текстової інформації;

– виконати програмну реалізацію розробленого алгоритму автоматичної класифікації текстової інформації;

– підготувати данні для оцінки якості класифікації;

– провести дослідження ефективності розробленої інформаційної технології.

Об’єктом дослідження є процес класифікації неструктурованих текстових масивів інформації.

Предметом дослідження є технології та методи інтелектуального аналізу текстової інформації.

Методами дослідження є методи машинного навчання, які базуються на методах text mining.

Наукова новизна отриманих результатів. Розроблено модифікований метод індексації на основі статистичного алгоритму Вітербі з підключенням бази граматичних правил зняття морфологічної омонімії.

Публікації. Результати досліджень опубліковані в журналі «Науковий огляд» [**Error! Reference source not found.**],опубліковані в тезах науково практичної конференції математичне та імітаційне моделювання систем. МОДС "2017" [**Error! Reference source not found.**], опубліковано в тезах наукової конференції студентів, магістрантів та аспірантів «Інформатика та обчислювальна техніка» – ІОТ-2018 [**Error! Reference source not found.**].

МАШИННЕ НАВЧАННЯ, TEXT MINING, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ТЕКСТОВИЙ КЛАСИФІКАТОР, АНАЛІЗ КОНТЕНТУ, АЛГОРИТМИ КЛАСИФІКАЦІЇ, ІНДЕКСАЦІЯ ТЕКСТІВ