**Місуно Іван Семенович. Розробка і дослідження векторних представлень інформації для задач пошуку і класифікації : Дис... канд. наук: 05.13.23 – 2006**

|  |  |
| --- | --- |
| |  | | --- | | **Місуно І.С. Розробка і дослідження векторних представлень інформації для задач пошуку і класифікації. – Рукопис.**  Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – Системи та засоби штучного інтелекту. – Інститут проблем математичних машин і систем НАН України, Київ, 2006.  Дисертація присвячена розвитку і підвищенню ефективності методів векторного представлення й обробки візуальної і текстової інформації в задачах пошуку і класифікації. Розроблено методи добору бінарних ознак, що засновані на використанні критеріїв інформативності і надмірності. Методи забезпечують у задачах класифікації зниження обчислювальних витрат при збереженні якості. Розроблено методи підвищення якості класифікації даних, що мають векторне представлення, шляхом вдосконалення правила навчання персептрона, цілеспрямованого розширення навчальної вибірки, об'єднання множинних результатів класифікації і застосування добору найбільш надійних результатів, а також відмови від класифікації. Розроблено методи формування розподілених представлень текстової інформації з урахуванням її семантичної близькості, що відрізняються застосуванням контекстних векторів з дискретними елементами і використанням текстів навчальної бази як контекстів. Ефективність всіх розроблених методів експериментально підтверджено дослідженнями на реальних даних. Розроблено засоби підтримки створення і реалізації інтелектуальних інформаційних технологій класифікації і пошуку, що використовують оригінальні методи формування й оперування векторними представленнями інформації. Розроблено програмний нейрокомп'ютер *SNC*, який є інструментальнім засобом для розробки, дослідження і практичного застосування методів інтелектуальної обробки даних. | |
| |  | | --- | | Сукупність отриманих у дисертації результатів забезпечує вирішення актуальної наукової задачі розвитку й підвищення ефективності методів векторного представлення й обробки інформації для задач пошуку і класифікації. Розроблені і досліджені методи представлення й оперування розрідженими векторними представленнями, відбору інформативних ознак, урахування семантичної близькості текстової інформації, класифікації і пошуку. Підвищена ефективність розв’язання задач пошуку і класифікації текстової і візуальної інформації, що показано експериментальними дослідженнями на реальних даних. Методи реалізовано у програмних засобах, які можуть бути використані як компоненти інтелектуальних інформаційних технологій.  За результатами проведеного дослідження зроблено такі висновки:  1. Розроблені методи добору бінарних інформативних ознак дозволяють зменшувати розмірність векторних представлень під час класифікації векторних даних. Методи, реалізовані програмними засобами *Feature Toolbox*, забезпечують зниження обчислювальних витрат при збереженні якості в задачах класифікації, що підтверджено на реальних даних. Так, для зображень бази *MNІST* при скороченні числа ознак у 10-50 разів отримана точність класифікації більше 97% і збільшена швидкість класифікації більше ніж у 10 разів; для текстів бази *Reuter-21578* при скороченні розмірності текстових векторів у 20-100 разів отримано результат 0,903–0,937, близький до кращих світових (0,92).  2. Розроблено методи, що забезпечують підвищення якості класифікації даних, які мають векторне представлення, шляхом удосконалення навчання, об'єднання множинних результатів класифікації, застосування добору найбільш надійних результатів і відмови від класифікації. Методи реалізовано програмними засобами *Classifier Toolbox*та в нейрокомп’ютері *SNC* і експериментально досліджені. На базі *MNIST* об'єднання множинних результатів класифікації і застосування відбору найбільш надійних результатів дало зменшення помилки класифікації з 0,72% до 0,60%, а застосування відмови від класифікації дозволило додатково знизити рівень помилок.  3. Розроблено методи формування розподілених представлень текстової інформації з урахуванням її семантичної близькості, що відрізняються застосуванням контекстних векторів з дискретними елементами. Методи реалізовано програмними засобами *Semantіc Toolbox* і використано у макеті програмної системи контекстного пошуку текстів *TextSearch*, реалізованої в веб-серверній архітектурі. Показано поліпшення результатів пошуку на стандартних базах текстів *Medlars*, *Cranfіeld*, *Tіme* до 20% за 11-точковою усередненою характеристикою порівняно з *VSM*-пошуком. При класифікації бази *Reuter*-21578 за допомогою *Semantic Toolbox* та *Classifier Toolbox* отримано результати до 0,918, що відповідає рівневі кращих результатів традиційних методів.  4. Розроблені методи формування й оперування векторними представленнями дозволяють пропорційно квадрату розрідженості векторів зменшувати витрати часу на виділення ознак, обчислення інформативності ознак, класифікації і пошуку інформації, завдяки використанню уніфікованого представлення інформації у форматі розріджених векторів і обробки тільки значимих елементів. Методи реалізовано в об’єктно-орієнтованій бібліотеці класів *С*++ *CommonLіb*, використання якої дозволяє підвищити обчислювальну ефективність інформаційних технологій обробки текстів і зображень.  5. Створений програмний нейрокомп’ютер *SNC* є ефективним інструментальним засобом розробки нейромережевих інтелектуальних ІТ. Ефективність забезпечується архітектурою *SNC*, яка підтримує візуальне конфігурування проектованих алгоритмів у вигляді блок-схем, дозволяє створювати гнучко конфігуровані проекти для дослідницьких і прикладних цілей, розширювати клас вирішуваних задач шляхом включення нових функціональних модулів.  6. Розроблено нові програмні засоби для створення і реалізації інтелектуальних ІТ класифікації та пошуку, що використовують оригінальні методи формування й оперування векторними представленнями інформації (*CommonLіb*, *Feature Toolbox*, *Classifier Toolbox*, *Semantіc Toolbox*), та макети, що демонструють їх ефективність (*TextSearch, TextClassifier*).  7. Результати проведених досліджень показали перспективність застосування розроблених методів формування й обробки векторних представлень, а також розроблених програмних засобів для створення нових інтелектуальних інформаційних технологій, що включають задачі пошуку і класифікації візуальної і текстової інформації. Практична значимість розробок підтверджується актами впровадження. | |