**Кабак Леонід Віталійович. Моделі та методи класифікації текстових документів в спеціалізованих інформаційно- пошукових системах : Дис... канд. наук: 05.13.06 - 2006.**

|  |  |
| --- | --- |
|

|  |
| --- |
| Кабак Леонід Виталійович. Методи та моделі класифікації текстових документів в спеціалізованих інформаційно-пошукових системах. – Рукопис.Дисертація на здобуття вченого ступеня кандидата технічних наук за спеціальністю 05.13.06 – автоматизовані системи управління та прогресивні інформаційні технології. – Харківський національний університет радіоелектроніки, Харків, 2006.Дисертація присвяченарозробці сучасних методів класифікації, які можна застосовувати під час обробки великих масивів текстової документації в системі документообігу Єдиної автоматизованої інформаційної системи (ЄАІС) митної служби України.Однимз актуальних для митної служби України є питання класифікації документів, які надходять до бази даних автоматизованої інформаційно-пошукової системи (АІПС), за допомогою поштового сервера внутрішньої митної мережі INTRANET. У дисертаційній роботі пропонуються методи автоматизованої класифікації документів, які враховують відносну частоту появи певних термінів зі словника даних (глосарія) у документах, які належать до різних тематичних розділів, а також взаємну залежність появи різних термінів у документах певного класу. |

 |
|

|  |
| --- |
| У дисертаційній роботі наведено результати, які, відповідно до мети дослідження, у сукупності є рішенням актуального наукового й практичного завдання класифікації текстових документів у спеціалізованих інформаційно-пошукових системах. Отримані методи розроблено з метою підвищення ефективності роботи інформаційно-пошукових систем, які використовуються в Єдиній автоматизованій інформаційній системі митної служби України. Вирішення цієї проблеми отримано на основі застосування розроблених моделей і методів обробки текстових документів в автоматизованих інформаційно-пошукових системах. З виконаних теоретичних та експериментальних досліджень можна зробити такі висновки:1. Досліджено існуючі методи класифікації текстових документів, які використовуються в інформаційно-пошукових системах, показано, що на сьогоднішній день розроблено й використовується досить велика кількість математичних моделей і чисельних методик класифікації документів. Однак розроблені моделі й методи алгоритмічно складні і не враховують особливості програмного забезпечення сервера СУБД Oracle.2. Удосконалено метод одиничної класифікації текстової документації, заснований на ймовірності входження певного терміна в документ заздалегідь визначеного класу, що дозволяє автоматизувати класифікацію документів, які надходять до системи. Цей метод відрізняється від існуючих гнучкістю та простотою настроювання параметрів класифікації.3. Уперше розроблено метод множинної класифікації текстових документів в інформаційно-пошукових системах, заснований на ймовірностях появи в документах груп термінів зі словника даних, які складаються з двох або більше термінів, що дає можливість отримувати похибки класифікації належного рівня та підвищити якість класифікації. Розроблений метод класифікації відрізняється від існуючих адаптованістю, можливістю варіювання параметрів таких, як кількість термінів у документі й урахування їх одночасної появи. Запропонований метод має можливість регулювання сукупного параметра “якість класифікації” й гарну пристосованість до практичної реалізації на сучасних реляційних СУБД.4. Набув подальшого розвитку метод вибору критерію якості класифікації, який, завдяки введенню вагових коефіцієнтів, дозволяє регулювати параметри “якість класифікації – витрати обчислювальних ресурсів”, що дає можливість підвищити ефективність системи автоматизованої класифікації текстових документів5. Розроблено метод формалізованого опису коефіцієнта довіри, завдяки якому з’явилась можливість зарахувати документ, що надходить до системи, як до одного, так і до декількох класів і використовувати у методі множинної класифікації, що підвищує якість класифікації документів.6. Набула подальшого розвитку модель системи класифікації текстових документів завдяки запропонованому критерію якості класифікації та розробленим методам, що дозволило отримувати похибки класифікації належного рівня з прийнятною витратою обчислювальних ресурсів системи. Модель та методи класифікації текстових документів дозволяють оптимально використати можливості сервера БД Oracle для підвищення продуктивності при автоматизованій класифікації, пошуку й обробки великих масивів текстових документів у рамках ЄАІС митної служби України.7. Отримані результати впроваджено як елемент ЄАІС митної служби України, що дозволило істотно підвищити якість функціонування інформаційно-пошукової системи ЄАІС митної служби України завдяки розвитку існуючих та розробці нових методів класифікації текстових документів, які забезпечують скорочення часу, зменшення обсягів щодо підготовки та пошуку даних в інформаційно-пошукових системах, про що свідчить акт впровадження Управління інформаційного забезпечення та митної статистики державної митної служби України (акт упровадження від 14.06.2004). Отримані результати впроваджено у вигляді програмного модуля для організації обробки й зберігання документів у відділі по боротьбі з контрабандою й порушеннями митних правил Дніпровської регіональної митниці (акт упровадження від 21.09.2000). Теоретичні результати дисертаційної роботи у вигляді моделей подання знань і методів обробки інформації впроваджені в навчальний процес Академії митної служби України на кафедрі інформаційних систем і технологій у дисципліні “Організація баз даних і знань” і використовуються в курсовому й дипломному проектуванні (акт упровадження від 25.10.2005 ). |

 |