

МЕТОД КЛАСІФІКАЦІЇ НЕСТРУКТУРОВАНИХ ДАНИХ В МУЛЬТИМЕДІА-СИСТЕМАХ

Святкін Я.В., Шубін І.Ю.

Харківський національний університет радіоелектроніки, Харків

Метою, що поставлена у дослідженні є репрезентація та класифікація неструктурованих даних в мультимедіа системах; запропоновано використання методу компараторної ідентифікації для розподілу на класи еквівалентності та зв'язування в гіперструктуру документів, що відібрані в результаті запиту до бази мультимедіа-даних (БМД).

Припустимо, що в результаті деякого запиту користувача до навчаючої системі із БМД були відібрані n документів $t_i, i = \overline{1, n}$. Усі мультимедіа-дані були отримані в результаті процедури обробки запиту, у відповідності з якою в індексних записах документів містяться одне або декілько понять $p_j, j = \overline{1, m}$ бази знань проблемної галузі (або потомків потрібних понять), що виконують роль аналога ключевих слів при індексуванні документів. Обозначимо множину документів БМД через $\mathbf{T} = \{t_i\}, 1 \leq i \leq n$; а множину вихідних понять проблемної галузі та відношень між ними – через $\mathbf{P} = \{p_j\}, 1 \leq j \leq m$.

На відмінність від традиційних систем, поняття, що розглядаються p_j – неоднородні об'єкти, які є елементами моделі проблемній галузі: поняття, властивості, значення цих властивостей, составні поняття та т.ін. Відносини між ними – довільні зв'язки. При цьому для представлення мультимедіа-даних з різних точок зору (зберігання та обробки, навігації і т.ін.), а також для реалізації різних стратегій або методик навчання може використовуватися множинне індексування. Таким чином, неоднорідність понять БЗ визначається також і різницею механізмів сгортання інформації БМД (представлення основного вмісту документів набіром ключових слів). В цілому же множина понять \mathbf{P} описує в стиснутому вигляді документи множини \mathbf{T} .

Тому для кожної пари t і p є точно визначенне значення відповідності (невідповідності) даного поняття проблемній галузі і документу, що розглядається. Ця відповідність може бути представлена предикатом, який назовемо предикатом релевантності R . Для кожної пари документу $t \in \mathbf{T}$ та поняття $p \in \mathbf{P}$ значення предикату дорівнює $\mathbf{1}$ у випадку відповідності даного поняття мультимедійному документу, та $\mathbf{0}$ – у протилежному випадку. При цьому важливо, щоб при кожному встановленню відповідності t та p предикат релевантності R определявся однозначно. У результаті при $R(t, p) = \mathbf{1}$, визначимо, що документ t є істинним відносно поняття p .