Вибір ознак в процесі інтелектуальної обробки текстових повідомлень

Марія Голуб

Кафедра журналістики, реклами та PR-технологій, Черкаський національний університет імені Богдана Хмельницького, УКРАЇНА, м. Черкаси, бульвар Шевченка 81, E-mail: Mas-golub@yandex.ru

Abstract – The process of selection of signs is investigational on results the decouplig of text messages. Certain features of the adaptive forming of informing signs are on the різниз levels of decouplig. Described process of intellectual analysis of properties of author of text message.

Ключові слова – текст, ознака, інформативність, моделювання, інтелектуальна обробка.

I. Вступ

Вже загальноприйнятим стало використання частотних характеристик лексем для класифікації текстових повідомлень [1]. Інтелектуальна обробка текстів зводиться до задач їх класифікації. Ми вважаємо, що це пов'язано із обреженою інформативністю частотних характеристик лексем. Разом з тим наші дослідження дозволяють стверджувати, що текстове повідомлення придатне для виявлення характеристик його автора, його психологічних станів, стану фізичного здоров'я, фаху та інше. Для цього необхідно виявити набір інформативних ознак та створити достатньо потужну інформаційну технологію. Дана роботи присвячена вибору достатнього переліку інформативних ознак, що здатні забезпечити дослідника відомостями про автора текстового повідомлення.

II. Результати досліджень

Була сформована гіпотеза про необхідність індивідуального визначення глибини декомпозиції тексту для кожної із поставлених задач інтелектуального аналізу текстового повідомлення. При цьому досліджувались рівні декомпозиції детальніші за лексему.

Для досліджень використовувались 3 авторські тексти тривалістю по три сторінки. Із кожного тексту форувались послідовність вибірок об'ємом 1000 знаків. Кожна вибірка описувалась частотними характеристиками елементів тексту, отриманими за результатом його декомпозиції. Досліджувалить масиви частотних характеристик для елементів тексту, отриманих за різною глибиною декомпозиції, що використовувались в якості ознак масиву вхідних даних (МВД). За кожним масивом ознак синтезувалась багатопараметрична індуктивна модель [2]. Інформативність набору ознак оцінювалась за

характеристиками результатів моделювання. Зокрема розраховувалось значення критерію регулярності [2].

Результатом досліджень є визначення глибини декомпозиції текстового повідомлення для кожної із поставлених задач: ідентифікації автора та визначення стану його фізичного здоров'я.

Дослідження містило кілька етапів. На першому етапі визначались перелік рівнів декомпозиції текстового повідомлення, задавався перелік ознак на кожному рівні, виявлялась їх повторюваність.

На другому етапі виявлялись ознаки, повторюваність яких є вищою заданого значення і на їх основі формувались МВД.

На третьому етапі синтезувалась модель, проводились її випробування на МВД, що не використовувались в процесі синтезу цієї моделі. Далі визначались характеристики результатів моделювання та робився висновок про інформативність ознак.

В результаті досліджень виявлено 144 інформативних ознак тексту, що забезпечують прийнятні характеристики результатів моделювання відповідно до поставленої задачі інтелектуальної обробки текстових повідомлень.

Висновок

Отримано експериментальне підтвердження гіпотези про індивідуальне визначення глибини декомпозиції текстового повідомленни для розв'язання різнотипних задач його інтелектуальної обробки. Адаптивність процесу формування інформативніх ознак забезпечується зміною глибини декомпозиції текстового повідомлення відповідно до особливостей поставлених задач.

Результати досліджень використовуються для створення інформаційної технології моніторингу авторів текстових повідомлень.

Література

2. Павлишенко Б. М. Формування базису семантичного простору текстових документів за допомогою генетичних алгоритмів / Б. М. Павлишенко // Математичні машини і системи. – 2013. – № 2. – С. 96-104.
3. Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем / А. Г. Ивахненко. – К. : Наук. думка, 1981. – 296 с.