Інтелектуальний аналіз текстів

Марія Голуб[1]

1. Кафедра інформаційної безпеки та комп'ютерної інженерії, Черкаський державний технологічний інститут, УКРАЇНА, м.Черкаси, бул Шевченка, 460,
E-mail: sanmarkovets@gmail.com

Author propose the Information technology of text mining. Experimentally confirmed the possibility of automating tests for a given search content. Number Field classified texts ranges from 65.4% to 100%.

Ключові слова – інтелектуальний аналіз, текст, інформаційний моніторинг, класифікація.

Інтелектуальний аналіз текстів має на меті синтезувати моделі, здатні виконати пошук інших текстів, що схожі за змістом. Необхідно розв'язати задачу класифікації ділянок текстів.

Для цього необхідно: 1) класифікувати ділянки тексту за змістом – виявити кількість змістових класів в заданому тексті та визначити розмір та кількість ділянок тексту, що можуть бути віднесені до кожного із класів; 2) визначити перелік класифікаційних ознак, що поєднують ділянки тексту; 3) оцінити можливість їх прямого застосування – встановити чи можливо застосувати вибрані ознаки для класифікації наступних ділянок тексту за їх граничними значеннями; 4) синтезувати правила класифікації нових ділянок за заданими ознаками. Таке правило може бути у вигляді граничних значень класифікаційних ознак або ж у вигляді аналітичного виразу, що перетворює значення множини ознак в результат класифікації – висновок про належність ділянки тексту до даного класу.

Дослідження показали, що явним чином не удається виявити ні перелік класифікаційних ознак, ні вирішуюче правило, що побудоване на основі цих ознак. Тому для кожного із класів виявляється свій набір ознак і на їх основі будується окреме вирішуюче правило, що дозволяє розпізнати ділянки тексту, що належать до заданого змістового класу.

Правила синтезувались у вигляді багаторівневих функціональних залежностей, що поєднують в своїй структурі індуктивні моделі, нейромережі, генетичні алгоритми, гібридні моделі, отримані за завершеними алгоритмами. Для розв'язання задачі узгодження взаємодії цих моделей використовується метод висхідного синтезу елементів [1]. Ці методи, засоби та алгоритми синтезу моделей (АСМ) поєднані в інформаційну технологію інтелектуального аналізу текстів (ІТІАТ). Вона містить такі етапи.

1. Формування вирішуючих правил. Формуються вирішуючі правила у вигляді багатошарових аналітичних моделей. Проводиться аналіз тексту, виявляється перелік інформативних ознак, що здатні розв'язати поставлену задачу. За вибраними ознаками забезпечується перетворення текстового повідомлення в масив чисельних характеристик інформативних ознак, синтез та випробування моделей.

Текст аналізується шляхом виділення окремих ділянок (вікон) із послідовно розміщених в тексті речень [2]. Перетворення текстового повідомлення

проводиться шляхом розрахунку частотних характеристик наперед заданих ознак тексту. В результаті розрахунків отримуємо вектор ознак — точку спостереження в багатовимірному просторі ознак первинного опису об'єкта моделювання. Після оцінки інформативності деякі ознаки, що мають незадовільні частотні характеристики, текстів видаляються із первинного опису. Точки спостереження, що сформовані за вектором інформативних ознак утворюють масив вхідних даних (МВД).

Належність ділянки тексту до заданого змістовного класу відображалась шляхом позначення точки спостереження як «Свій». Таким чином утворюється послідовність точок спостереження, які відображають в своїй структурі зміст повідомлення, що належить до заданого класу. Іншим точкам спостереження надавався статус «Чужий». Вони відображають зміст повідомлення, які належать до інших класів.

Синтез моделей забезпечує розв'язок задачі групування точок спостереження. Кожна модель використовується як вирішуючи правило, що дозволяє віднести кожну точку до одного із заданих класів. Для формування цього правила забезпечується ієрархічне поєднання багатошарових моделей [1] відповідно до індуктивного методу [3]

За результатами випробування моделі визначалося порогове значення результатів моделювання, вище якого модельоване спостереження позначалось як «Свій». Після цього модель заноситься до бази модельних знань.

2. Класифікація текстів. Аналізований текст подається на вхід ІТІАТ. За результатами перетворення МВД паралельно кількома вирішуючими правилами бази модельних знань приймається рішення про відповідність дослідженого фрагменту тексту заданому змісту.

При випробуванні описаної технології визначалась кількість вірно розпізнаних фрагментів тексту відповідно до наперед заданого змісту та залежність результатів контекстного пошуку від виду опорної моделі. В залежності від вигляду опорної моделі кількість вірно класифікованих точок спостереження змінювалось в межах від 65,4% до 100,0%.

Таким чином запропоновано новий метод інтелектуального аналізу текстів, що дозволя автоматизувати процеси пошуку текстових повідомлень за змістом. Таким чином забезпечується можливість організувати інтелектуальну систему багаторівневого інформаційного моніторингу.

Література

1. Голуб С.В. Багаторівневе моделювання в технологіях моніторингу оточуючого середовища / С.В. Голуб. – Черкаси: Вид. від. ЧНУ імені Богдана Хмельницького, 2007. – 220 с 2. Голуб С.В. Відображення властивостей автора тексту в структурі багатопараметричної моделі/ С.В. Голуб, О.В. Константиновська, М.С. Голуб// Системи обробки інформації: Збірник наукових праць. – Х.: Харківський університет повітряних сил імені Івана Кожедуба, 2014. – Вип. 9 (125). – С. 82-8. 3. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем / А.Г. Ивахненко. – К. : Наук. думка, 1981. – 296 с.