ПІДХІД ДО КЛАСИФІКАЦІЇ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ, ЩО НАДХОДЯТЬ ЧЕРЕЗ RSS-КАНАЛИ

Ямшанов І.С.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

RSS (Rich Site Summary) є назвою родини XML-подібних форматів, що використовуються у мережі Internetдля публікації оновлень для сайтів. Через такі оновлення надається інформація про новини, статі, записи у блогах та подібні публікації, що з'являються на сайтах.

Для отримання таких оновлень використовуються RSS агрегатори – десктопні або веб-застосунки, що автоматично перевіряють появу нових публікацій для обраних сайтів. Пересічний користувач отримує оновлення від декількох десятків сайтів, що в середньому може становити сто та більше повідомлень за добу. Ознайомлення з такою кількістю інформації може зажадати значної кількості часу, що не завжди є прийнятним. Зрозуміло, що повідомлення мають різну цінність та рівень цікавості для користувача, і в першу чергу він хотів би ознайомитись з найбільш цікавими. Варто відзначити, що агрегатори зазвичай не реалізують функціональності для сортування та відбору записів на основі тих чи інших критеріїв, окрім можливості вказати категорію, до якої будуть належати усі публікації від конкретного сайту.

Тож існує проблема визначення для кожного з вхідних повідомлень рівня його цікавості для користувача, та потреба реалізації відображення повідомлень у агрегатові відповідно до цього рівня.

Визначення цікавості має базуватись на наступних засадах:

- рівень цікавості повідомлень є індивідуальним для користувачів;
- критерії визначення цікавості для користувача можуть відрізнятись для різних сайтів, з яких він отримує оновлення;
- з часом рівень цікавість конкретних тем може змінюватись.

Пропонується наступний підхід до класифікації повідомлень:

1) Класифікація відбувається з використанням нейронних мереж. На вхід мережі надходить інформація про повідомлення: належність до категорії, довжина повідомлення, співвідношення текстового та графічного вмісту повідомлення, час, що минув з моменту публікації повідомлення, результат семантичного аналізу тексту повідомлення за допомогою класифікатора Баєса, тощо. На виході мережі буде оцінка цікавості повідомлення на інтервалі [0;1].

2) На початку класифікації використовується єдина мережа для класифікації усіх повідомлень. Надалі, коли кількість повідомлень у окремій категорії або для окремого сайту досягне деякого порогового значення, нова нейронна мережа буде створюватись, навчатись та використовуватись для класифікації у цій категорії або для цього сайту, що дозволить збільшити точність класифікації.

3) Повідомлення відображаються відповідно до зменшення рівня цікавості.

Архітектура нейронної мережі та критерії повідомлень, що використовуються для класифікації, є предметом подальших досліджень.