КЛАСИФІКАЦІЯ ТЕКСТІВ ЕКОНОМІЧНОГО НАПРЯМКУ НА ОСНОВІ СЕМАНТИЧНИХ МЕРЕЖ ГІПОНІМІЧНИХ ВІДНОШЕНЬ

Булатнікова Т.С.
Національний Технічний Університет
«Харківський Політехнічний Інститут»
м. Харків, вул. Пушкінська, 79/2, тел. 0995692641,
e-mail: tbu2641@gmail.com

Семантична мережа – це інформаційна модель предметної області, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної області, а ребра задають відношення між ними. Об'єктами можуть бути поняття, події, властивості та процеси.

Семантичні мережі відносяться до моделей класичного представлення знань у задачах штучного інтелекту, навчальних системах, системах машинного перекладу та семантичних павутинах. Підхід семантичних мереж базується на трьох основних складових: суб'єкт, відношення, об'єкт. Саме ці три складові є базовими блоками семантичних мереж.

Мережеві моделі формально можна задати у вигляді H = <I, C1, C2,..., Cn, R>, де I – множина інформаційних одиниць; C1, C2,..., Cn, – множина типів зв'язків між інформаційними одиницями. Відображення R задає між інформаційними одиницями, що входять до I, зв'язки із заданого набору типів зв'язків [1].

Гіпонімія як родо-видове відношення – це сукупність семантично однорідних одиниць, які належать до одного класу. На основі гіпонімії лексичні одиниці об'єднуються в тематичні й лексико-семантичні групи і поля.

Сьогодні гіпонімічні відношення широко використовуються для класифікації компаній за галуззю виробництва. Класифікація за галуззю виробництва впорядковує компанії у виробничі групи, які основані на схожості способів виробництва, продукції або поведінки на фінансовому ринку. Такі угрупування широко використовуються статистичними агенціями у фінансовій сфері послуг для групування схожих інвестиційних компаній при створенні індексів фінансового ринку за секторами [2].

Класифікація текстів – одне із завдань інформаційного пошуку, яке полягає у віднесенні тексту до однієї з кількох категорій.

Класифікація може здійснюватися повністю вручну, автоматизовано за допомогою створеного вручну набору правил або автоматично із застосуванням методів машинного навчання. До автоматичних методів класифікації текстів належать:

- EM-алгоритм;
- наївний байєсівський класифікатор;
- tf-idf;
- латентно-семантичний аналіз;
- штучна нейронна мережа;

– метод k найближчих сусідів;
– дерево прийняття рішень;
– тощо[3].

Класифікація текстів на основі семантичної мережі полягає в тому, що ключові слова, за допомогою яких визначатиметься приналежність тексту до певної теми, та самі теми знаходяться в безпосередньому зв'язку одне з одним. Така класифікація дозволить розширити можливості систем автоматичної обробки англомовних текстів економічної тематики, інформаційного пошуку, систем для вилучення інформації, систем аналізу тональності текстів, систем вилучення інформації, систем автоматичного реферування тощо.

Розроблена семантична мережа зберігається у форматі JSON. Такий формат є універсальним ще й тому, що майже всі сучасні мови програмування підтримують роботу з даними, які зберігаються в цьому форматі.

В якості вхідних даних програма бере обраний користувачем файл, в якому збережена потрібна кількість статей новин економічної тематики або вводить їх вручну у відповідне поле. Після того, як програма отримала текстові дані виконується наступний алгоритм:

1. Текст розбивається на статті відповідно до вказаного роздільника та створюється список.

2. Виконується пошук на виявлення економічних понять, якщо таке поняття було знайдене то стаття заноситься в новий список разом із співвіднесеним поняттям, його підгрупою та групою, інакше стаття заноситься в інший список.

3. Виконується перевірка на наявність однієї і тієї ж статті з різними поняттями, якщо такі знайдено, то виконується їх групування.

4. Кожна стаття перевіряється на наявність назви компанії, якщо компанію було знайдено то до статті «прикріплюється» назва компанії та місце її розташування.

5. Якщо компанію не знайдено, то стаття перевіряється на наявність країни, якщо країну знайдено, то статті присвоюється назва країни.

6. Якщо в статті не було знайдено ні країни ні компанії, то стаття заноситься в загальний список без них.

7. В результаті з'являється нове вікно з класифікованими статтями.

Розроблене програмне забезпечення класифікує статті економічного напрямку за тематиками, разом із тематикою зі статті вилучається компанія або країна, назва якої зустрічається в статті, а також для компанії визначається місце розташування її головного офісу. Результат роботи програми користувач може зберегти у файлі, відредагувати або просто продивитися.

Список литературы

1. Roussopoulos N.D. A semantic network model of data bases. – Department of Computer Science, University of Toronto, 1976. – p. 104.

2. Day, A.C.L. The taxonomic approach to the study of economic policies. – The American Economic Review, 2010. – p. 78

3. Hastie, T. The Elements of Statistical Learning. Springer, 2001. – C.758