А. А. Яровий
Д. С. Кудрявцев

МЕТОД СИНХРОНІЗАЦІЇ ДАНИХ ТЕРМІНОЛОГІЧНИХ БАЗ ЗНАНЬ

Вінницький національний технічний університет

Запропоновано метод синхронізації даних в термінологічних базах знань на основі використання нейронної мережі та класифікації за тематикою предметних областей. Синхронізація текстових даних є однією з ключових задач для ефективної обробки даних, що полягає у систематизації знань за семантичною складовою та вирішує проблему розуміння контексту на основі вже відомих даних. Для вирішення даної задачі використовують комплексний підхід, що включає в себе набір рішень та алгоритмів синхронізації на усіх рівнях обробки даних, включаючи семантичний аналіз, алгоритми програмного та апаратного рівнів, а також використання оптимізованих моделей даних для конкретних задач. В ході дослідження розроблено алгоритм пошуку оптимального розподілу даних в термінологічних базах знань на основі семантичної цінності термів. Визначено основні критерії розподілу текстових даних в термінологічних базах знань. Розглянуто задачу актуалізації даних в термінологічних базах знань при їх наповненні. Сформульовано та описано задачу конфліктної синхронізації термів на основі семантичної належності до споріднених предметних областей. Розглянуто алгоритми порівняння термів на основі семантичного аналізу, косинусної подібності, коефіцієнту Жаккара та методу частоти появи термів (TF-IDF). Подано опис методу синхронізації у відповідності до створених моделей та структур даних. Описано переваги та недоліки відомих методів синхронізації текстових даних із використанням алгоритмів та методів обробки даних на прикладі задач збереження та відтворення даних. Відзначено ефективність методу синхронізації в ході тестування та експериментальних досліджень для кол-центрів. Оптимізовано структури даних для ефективного збереження та отримання текстових даних для задачі класифікації тексту. Створено прототип термінологічних баз знань та застосовано метод синхронізації на прикладі різних предметних областей.

Ключові слова: термінологічна база знань, синхронізація даних, класифікація тексту, нейронна мережа, семантичний аналіз тексту.

Вступ

Синхронізація даних у термінологічних базах знань є однією з основних проблем, з якими стикаються дослідники та розробники, які працюють з великими масивами термінів та їх визначень [1]. Термінологічні бази знань (ТБЗ) є ключовим елементом для багатьох галузей науки і техніки, де важливо мати чітку та узгоджену термінологію для ефективної обробки та обміну інформацією. З огляду на збільшення обсягів даних та зростання складності інформаційних систем, виникає необхідність не тільки в зберіганні даних, а й в їхній актуалізації, систематизації та підтримці на належному рівні цілісності. Однією з основних проблем у контексті синхронізації даних є семантична неоднорідність термінів. Це означає, що терміни, навіть якщо вони схожі за змістом, можуть бути описані різними словами, що ускладнює процес пошуку та оновлення інформації. Крім того, бази знань часто збагачуються новими даними, що потребує інтеграції нових термінів з існуючими та вирішення конфліктів між різними джерелами [1]. У цьому контексті виникає питання про ефективні методи синхронізації даних у ТБЗ, які б дозволяли автоматично оновлювати базу знань, інтегруючи нові дані без втрати їхньої якості та цілісності. Сучасні дослідження в області обробки даних і штучного інтелекту надають нові можливості для розв'язання цієї проблеми. Зокрема, використання нейронних мереж та методів класифікації дозволяє автоматизувати процес обробки великих обсягів інформації, забезпечуючи точність і гнучкість при роботі з даними різних предметних областей [2]. Метод синхронізації, запропонований у цій роботі, заснований на комплексному підході до обробки текстових даних, що включає в себе використання алгоритмів семантичного аналізу та нейронних мереж для автоматичного порівняння та класифікації термінів. Основною метою цього методу є підвищення ефективності роботи з термінологічними базами знань за рахунок

оптимізації процесів зберігання, оновлення та пошуку даних. Основна складність, яка виникає при розробці методів синхронізації для ТБЗ, полягає в забезпеченні семантичної узгодженості термінів. Термінологія, яка використовується в різних предметних областях, може бути значною мірою взаємопов'язаною, але водночас мати суттєві відмінності у значеннях. Тому розробка методу синхронізації повинна враховувати ці особливості, забезпечуючи як правильну класифікацію термінів, так і їх узгодженість між різними областями, що є ключовою задачею даного дослідження. Для вирішення цього завдання запропоновано алгоритм синхронізації, що враховує семантичну цінність термінів. Семантична цінність визначається на основі аналізу контексту використання термінів та їх частотної характеристики в конкретних предметних областях. Використання нейронної мережі дозволяє автоматизувати процес класифікації та порівняння термінів на основі їх семантичних ознак, що значно підвищує точність синхронізації даних. Важливим аспектом методу є також використання кількох алгоритмів для порівняння термінів, включаючи методи косинусної подібності [3], алгоритми на основі частоти появи термів (TF-IDF) [4] та використання коефіцієнту Жаккара [5]. Ці методи дозволяють порівнювати терміни за різними критеріями, такими як їх фонетична та семантична подібність, що сприяє точнішій ідентифікації схожих термінів навіть у випадку їх часткових відмінностей у написанні або використанні. Дослідження також розглядає проблему конфліктної синхронізації термінів, яка виникає при злитті даних з різних джерел. Конфлікти можуть виникати через різне тлумачення одного й того ж терміну в різних областях знань, що потребує спеціальних механізмів для вирішення таких ситуацій. Запропонований метод передбачає використання семантичного аналізу для виявлення і вирішення конфліктів на основі аналізу контексту та визначення пріоритетності джерел даних.

Метою роботи є розроблення методу синхронізації даних термінологічних баз знань на основі визначення семантичної цінності (Terminology Knowledge Base Data Sync (TKBDS)).

Результати дослідження

Існуючі методи синхронізації, що застосовуються у звичайних базах даних, переважно спрямовані на узгодження структурованих даних. У цих методах використовуються такі підходи, як транзакції, контроль версій та періодичне оновлення. Ці підходи добре працюють у випадках з даними, що легко піддаються структуруванню, але коли мова йде про текстові дані, ці методи стають неефективними. Термінологічні бази знань мають справу з незліченними варіантами написання та інтерпретації термінів, що потребує нових, більш гнучких підходів для порівняння і систематизації даних. Однією з найбільших проблем при роботі з термінологічними базами є семантична неоднорідність даних. Це означає, що один і той самий термін може мати різні значення залежно від контексту, в якому він використовується. Це пояснюється багатозначністю термінів, синонімією, різними варіантами написання та контекстуальною залежністю. Така проблема ускладнює процес пошуку та порівняння термінів, що вимагає залучення нових методів для семантичного аналізу тексту. Відомі підходи до порівняння текстових даних, такі як алгоритми на основі відстані Левенштейна або алгоритми порівняння за частотою появи термінів, не завжди дозволяють досягти високої точності [6]. Ці алгоритми, хоча й є корисними для пошуку схожих термінів, не можуть повністю врахувати семантичну різницю між ними, що часто призводить до помилкових результатів під час синхронізації. Проте, у сучасних дослідженнях все більшої популярності набувають підходи на основі нейронних мереж, які дозволяють вирішувати завдання синхронізації з урахуванням семантики тексту. Використання нейронних мереж у цьому контексті забезпечує новий рівень автоматизації процесу порівняння та систематизації термінів. Нейронні мережі здатні обробляти великі масиви текстової інформації, зокрема векторизувати терміни та аналізувати їх з точки зору семантичного змісту. Запропонований метод синхронізації базується на поєднанні рекурентної нейронної мережі з алгоритмами семантичного аналізу, що дозволяє

автоматизувати порівняння термінів та їх класифікацію за тематикою предметних областей [7]. Архітектура нейронної мережі, розробленої для цієї задачі, має багаторівневу структуру, що дозволяє виконувати деталізований аналіз тексту. На початковому рівні нейронна мережа перетворює текстові дані в вектори, які можуть бути оброблені далі. Потім ці вектори проходять через кілька шарів нейронної мережі, що дозволяє класифікувати терміни за їхніми семантичними ознаками. Останній шар нейронної мережі відповідає за кількість класів на які класифікують терми. Важливо, що така архітектура дозволяє не тільки визначити схожі терміни, але й зрозуміти їхнє семантичне значення у різних контекстах. Одним із головних елементів запропонованого методу є класифікація термінів за тематикою предметних областей. Це дозволяє забезпечити синхронізацію не тільки за формальними ознаками термінів, але й за їхнім змістом. Наприклад, терміни, що використовуються в медицині, можуть бути схожі на терміни, що використовуються в інженерії, але мати різні значення. Завдяки класифікації за тематикою предметних областей можливо чітко визначити, в якому контексті використовується той чи інший термін, що допомагає уникнути помилок під час синхронізації. У контексті запропонованого методу використовуються кілька алгоритмів для порівняння термінів. Одним із найпопулярніших є алгоритм косинусної подібності, який дозволяє оцінити подібність між двома текстовими рядками на основі їхніх векторних представлень [3]. Алгоритм косинусної подібності обчислює кут між векторами двох текстів у багатовимірному просторі. Якщо кут малий, це означає, що тексти схожі, і навпаки, великий кут вказує на велику різницю між текстами. Це дозволяє точно визначити схожі терміни, навіть якщо вони мають різні слова чи структуру речень. Ще одним важливим елементом, що використовується в методі, є використання коефіцієнту Жаккара [5]. Він вимірює схожість між двома множинами елементів, порівнюючи відношення перетину цих множин до їхнього об'єднання. У контексті порівняння термінів це дозволяє оцінити, наскільки схожі терміни з точки зору їхніх складових частин. Коефіцієнт Жаккара корисний для знаходження термінів, які можуть бути описані різними словами, але мають схожі ознаки. Також у методі використовується алгоритм TF-IDF (Term Frequency-Inverse Document Frequency), що є одним із найпоширеніших методів для визначення важливості термінів у тексті. Цей алгоритм враховує частоту появи терміну у конкретному тексті та інверсну частоту його появи у всьому корпусі документів [4]. Це дозволяє точно визначати ключові терміни для кожної предметної області та використовувати їх для класифікації. Один із важливих аспектів запропонованого методу — це розробка алгоритму для пошуку оптимального розподілу даних у термінологічних базах знань. Визначення семантичної цінності термінів дозволяє ефективно класифікувати терміни та розподіляти їх у базі знань на основі їхнього контекстуального значення. Для цього було розроблено алгоритм, який автоматично аналізує частоту появи термінів у різних текстах і визначає їх семантичну важливість для конкретної предметної області. Основними критеріями для оптимального розподілу даних є частотність появи термінів, їхня семантична цінність та відповідність певним предметним областям. Запропонований алгоритм дозволяє автоматично класифікувати терміни на основі цих критеріїв, забезпечуючи їх правильний розподіл у базі знань. Базову схему алгоритму наведено на рисунку 1. Це дозволяє не тільки полегшити процес пошуку даних, але й забезпечити їхню актуальність і точність у кожній з предметних областей. Окрім цього,

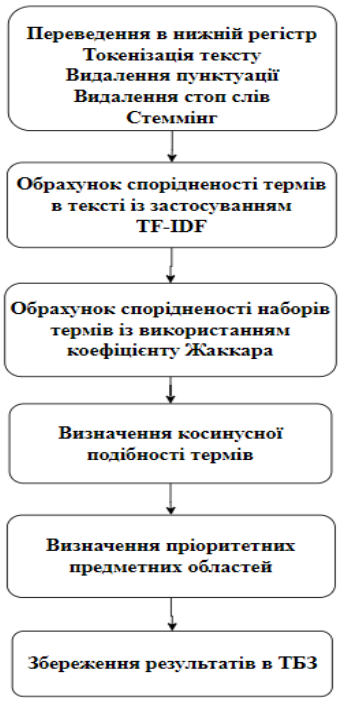

Рис. 1. Схема алгоритму синхронізації термів в ТБЗ

важливим аспектом методу є вирішення проблеми конфліктної синхронізації даних. Конфлікти виникають тоді, коли різні джерела надають суперечливі дані щодо одного й того ж терміну. Це може бути результатом різних інтерпретацій термінів у різних предметних областях або через використання різних варіантів написання одного й того ж терміну. Для вирішення цих конфліктів запропонований метод використовує семантичний аналіз, який дозволяє визначити ступінь спорідненості термів між собою та обирати найбільш релевантне джерело для оновлення бази знань. У рамках методу також було розроблено стратегії вирішення конфліктів на основі пріоритетності джерел інформації. Це дозволяє автоматично визначати, які джерела є найбільш надійними та актуальними для певної предметної області, і використовувати ці джерела для злиття даних. Таким чином, можна забезпечити узгодженість даних у базі знань і уникнути помилкових оновлень.

Висновки

Метод синхронізації був протестований на прикладі термінологічної бази знань, яка використовується у кол-центрі [8]. Кол-центри, як правило, оперують великою кількістю текстової інформації, що включає клієнтські запити, відповіді операторів та інші дані, пов'язані з обслуговуванням клієнтів. Термінологічна база знань для кол-центру допомагає класифікувати і зберігати всю інформацію про клієнтську підтримку, що є важливим для швидкого доступу до інформації та ефективної роботи операторів. У ході дослідження було створено прототип термінологічної бази знань для кол-центру, в якому було реалізовано запропонований метод синхронізації. Прототип дозволив автоматизувати процес зберігання та обробки інформації, значно скоротивши час на пошук потрібних даних.

Таблиця 1 – Результати тестування методу синхронізації.

Предметна область	Телефонна компанія	Служба підтримки користувачів побутової техніки	Служба гарантійного сервісу СТО
Початкова кількість термів	4283	6909	5861
Кількість нових термів	552	285	165
Загальна кількість текстів	1042	2639	1215
Середня довжина тексту, символів	448,37	387,04	323,73
Можливість виникнення конфлікту терму між предметними областями на 1000 термів (Word2Vec), %	5,31%	3,01%	2,97%
Можливість виникнення конфлікту терму між предметними областями на 1000 термів (TF-IDF), %	3,76%	2,05%	2,11%
Можливість виникнення конфлікту терму між предметними областями на 1000 термів (TKBDS), %	**3,54%**	**1,74%**	**2,05%**
Час синхронізації Word2Vec, с	0,327	0,424	0,329
Час синхронізації TF-IDF, с	0,417	0,512	0,385
Час синхронізації TKBDS, с	**0,299**	**0,391**	**0,284**

Результати впровадження показали, що метод синхронізації дозволяє підвищити ефективність роботи з термінологічними даними, забезпечуючи більш швидкий доступ до необхідної інформації на 7,6-8,6% у порівнянні з іншими методами. Оцінка ефективності показала, що застосування нейронних мереж та алгоритмів семантичного аналізу дозволяє суттєво підвищити ефективність синхронізації даних у великих текстових масивах.

Запропонований метод дозволив дещо знизити кількість помилок при пошуку та обробці інформації, що є критично важливим для систем, які працюють з великими обсягами даних на 3-5,9% у порівнянні з методом TF-IDF. Крім того, оптимізовані структури даних, розроблені в рамках цього дослідження, забезпечили ефективне збереження інформації та її швидке відновлення, що є особливо важливим для задач класифікації та зберігання текстових даних. Запропонований метод синхронізації даних у термінологічних базах знань є ефективним інструментом для автоматизації обробки великих обсягів інформації. Використання рекурентної нейронної мережі та алгоритмів семантичного аналізу дозволяє не тільки забезпечити точну синхронізацію, але й врахувати семантичну цінність термінів, що дозволяє досягти високої точності у класифікації даних між різними предметними областями. В подальшому планується детальний огляд даного методу синхронізації в термінологічних базах знань з описом усіх етапів та проведення експериментального дослідження на більшій кількості предметних областей.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] [Kaya, Cem & Kilimci, Zeynep & Uysal, Mitat & Kaya, Murat. (2024). A Review of Metaheuristic Optimization Techniques in Text Classification. International Journal of Computational and Experimental Science and Engineering. 10. 10.22399/ijcesen.295.

[2] Mohabir, S.E., Joshi, Y.C. A bibliometric analysis of the knowledge base on multinational corporations' behavior. *SN Bus Econ* 4, 105 (2024). https://doi.org/10.1007/s43546-024-00705-7.

[3] Ünver, Mehmet. (2023). Improved cosine similarity measures for q-Rung orthopair fuzzy sets. Qeios. 10.32388/EOGFR4.

[4] TF–IDF. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_832

[5] Travieso, Gonzalo & Benatti, Alexandre & da F. Costa, Luciano. (2024). An Analytical Approach to the Jaccard Similarity Index. 10.13140/RG.2.2.23119.70562.

[6] Berger, Bonnie & Waterman, Michael & Yu, Yun. (2020). Levenshtein Distance, Sequence Comparison and Biological Database Search. IEEE Transactions on Information Theory. PP. 1-1. 10.1109/TIT.2020.2996543.

[7] A. Yarovyi and D. Kudriavtsev, "Multi-purpose search to determine the context of a text message based on the dictionary data structure," *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, LVIV, Ukraine, 2021, pp. 65-68, doi: 10.1109/CSIT52700.2021.9648803.

[8] Gabriel A. (2020, January). Kensho Derived Wikimedia Dataset. Retrieved September 1, 2024 from https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data.

Яровий Андрій Анатолійович — д-р техн. наук, професор, завідувач кафедри комп'ютерних наук, e-mail: a.yarovyy@vntu.edu.ua;
Кудрявцев Дмитро Станіславович — аспірант кафедри комп'ютерних наук, e-mail: dmytro_k@vntu.edu.ua.

A. A. Yarovyi
D. S. Kudriavtsev

Method of terminological knowledge bases data synchronization

Vinnytsia National Technical University

Method for data synchronization in terminological knowledge bases is proposed, based on the use of a neural network and classification by subject area topics. Text data synchronization is one of the key tasks for efficient data processing, which involves systematizing knowledge based on its semantic component and addresses the problem of understanding context based on already known data. To solve this problem, a comprehensive approach is used, which includes a set of solutions and synchronization algorithms at all levels of data processing, including semantic analysis, software and hardware-level algorithms, and the use of optimized data models for specific tasks. During the research, an algorithm for optimal data distribution in terminological knowledge bases was developed based on the semantic value of terms. The main criteria for the distribution of textual data in terminological knowledge bases were identified. The task of updating data in terminological knowledge bases during their population was examined. The problem of conflicting term synchronization based on semantic affiliation to related subject areas was formulated and described. Algorithms for term comparison based on semantic analysis, cosine similarity, the Jaccard method, and the term frequency-inverse document frequency (TF-IDF) method were considered. A description of the synchronization method in accordance with the created models and data structures is provided. The advantages and disadvantages of known methods for text data synchronization using data processing algorithms and methods are described, with examples given of data storage and retrieval tasks. The effectiveness of the synchronization method is demonstrated using examples. Data structures were optimized for efficient storage and retrieval of text data for text classification tasks. A prototype of terminological knowledge bases was created, and the synchronization method was applied using the example of the call center subject area.

Keywords: terminological knowledge base, data synchronization, text classification, neural network, semantic text analysis.

Yarovyi Andrii Anatoliyovych — Professor, Head of the Computer Science Department, e-mail: a.yarovyy@vntu.edu.ua;

Kudriavtsev Dmytro Stanislavovych — Post-graduated student of the Computer Science Department, e-mail: dmytro_k@vntu.edu.ua