Запропоновано алгоритм пошуку подібності складно-структурованих даних із використанням семантичних мереж. Досліджено використання алгоритму для задачі пошуку подібності бібліографічних описів в інформаційно-аналітичній системі "ScienceLP". Для універсальності програмної реалізації алгоритму запропоновано використати рефлексивно-орієнтований підхід програмування

Ключові слова: алгоритм пошуку, бібліографічний опис, пошук подібності, семантичні мережі

Предложен алгоритм поиска сходства сложно-структурированных данных с использованием семантических сетей. Исследовано использование алгоритма для задачи поиска сходства библиографических описаний в информационно-аналитической системе "ScienceLP". Для универсальности программной реализации алгоритма предложено использовать рефлексивно-ориентированный подход программирования

Ключевые слова: алгоритм поиска, библиографическое описание, поиск сходства, семантические сети

УДК 004.822+004.912
DOI: 10.15587/1729-4061.2015.51051

АЛГОРИТМ ІДЕНТИФІКАЦІЇ ПОДІБНОСТІ СКЛАДНО-СТРУКТУРОВАНИХ ДАНИХ НА ОСНОВІ СЕМАНТИЧНИХ МЕРЕЖ

Р. Б. Тушницький
Кандидат технічних наук, доцент*
E-mail: ruslan.tushnytskyy@gmail.com
В. М. Макар
Кандидат технічних наук, доцент*
E-mail: makvm_cad@yahoo.com
*Кафедра програмного забезпечення
Національний університет «Львівська політехніка»
вул. С. Бандери, 12, м. Львів, Україна, 79013

1. Вступ

Однією з основних проблем обробки інформації є її інтелектуальний аналіз. Більшість існуючих методів в основному забезпечують обробку текстової інформації. Поряд з тим слабо досліджена методика інтелектуального аналізу складно-структурованих даних. На сьогодні однією із важливих задач є визначення подібності таких даних. Актуальність задачі ідентифікації подібності виражається у необхідності застосувати ці методи у різних сферах діяльності: пошук інформації; навчання; визначення плагіатів; системи колективної роботи; версіонування даних.

У більшості областей застосування вирішальними критеріями при виборі технологій і алгоритмів є швидкість роботи і гарантії забезпечення якості. Для забезпечення швидкодії доводиться відмовлятися від використання систем штучного інтелекту заснованих на базах знань, так як вони не витримують величезного потоку несистематизованої інформації різної тематичної спрямованості. Формування методик і засобів з оцінки якості порівняння інформації є відкритою проблемою для дослідження.

2. Аналіз літературних даних та постановка проблеми

На сьогодні існує ряд підходів до визначення подібності текстів. В методі *шинглів* для всіх ланцюжків аналізованого тексту розраховується так звана "сигнатура" – унікальне число, поставлене у відповідність деякому тексту і/або функція його обчислення [1]. Даний метод є досить ресурсоємний і його можна обійти, незначно змінивши текст, так як, насамперед, шингли залежать від відстані між словами.

Існуючі методи обчислення "сигнатур" поділяються на:

– *синтаксичні методи* – оперують з ланцюжками слів;

– *лексичні методи* – оперують зі словником.

Неважко показати, що за шингли можна з високою ймовірністю судити про подібність текстів, їх вкладеності, плагіаті і т. д. Однак для практичних завдань, в тому числі для виявлення масових розсилок, потрібно занадто велика кількість шинглів, що представляє високі вимоги до ресурсів для проведення процедури кластеризації.

Серед лексичних методів поширеним є *Match* або метод *«Описових слів»* [2]. Побудова контрольних сум для обмеженого числа (40–60 %) слів, які найбільш повно описують вміст тексту. Описові слова підбираються з урахуванням важливості слова. Важливість може визначатися динамічно або заздалегідь бути розрахована на тестової вибірці для оптимізації.

Відстань редагування. Вперше була визначена Левенштейном. Визначення можна поширити і для текстів, представляючи абзаци або речення як слова, а слова як символи. Раніше були розроблені різні реалізації алгоритмів, які найчастіше використовувалися для побудови систем перевірки орфографії, всі вони можуть бути адаптовані для визначення схожості текстів.

Відстань Левенштейна – це мінімальна кількість операцій вставки одного символу, видалення одного символу та заміни одного символу на інший, необхідних для перетворення одного рядка в інший. Відстань

Левенштейна та його узагальнення активно застосовується:

– для виправлення помилок в слові (в пошукових системах, базах даних, при введенні тексту, при автоматичному розпізнаванні відсканованого тексту або мовлення);

– для порівняння текстових файлів утилітою diff і їй подібними. Тут роль «символів» грають рядки, а роль «рядків» – файли;

– в біоінформатиці для порівняння генів хромосом і білків.

З точки зору додатків визначення відстані між словами або текстовими полями за алгоритмом Левенштейна, можна виділити наступні недоліки:

– при перестановці місцями слів або частин слів виходять порівняно великі відстані;

– відстані між абсолютно різними короткими словами виявляються невеликими, в той час як відстані між дуже схожими довгими словами виявляються значними.

Відстань Дамерау-Левенштейна – це міра різниці двох рядків символів, обумовлена як мінімальна кількість операцій вставки, видалення, заміни та перестановки сусідніх символів, необхідних для перекладу одного рядка в іншу. Є модифікацією відстані Левенштейна, відрізняється від нього додаванням операції перестановки.

Для задачі пошуку також відомі спроби поєднання лексичних і структурних мір подібності [3].

Лінгвістичні методи. Суть методів полягає в побудові дерева вмісту документу і його глибокому аналізі. Для підвищення якості роботи алгоритмів додатково може здійснюватися попередня обробка вхідної інформації. Найбільш ефективними є такі засоби підвищення якості [4–6]:

– *стеммінг* – нормування слова, приведення до єдиного кореня;

– *лінгвістичні бази* – бази перекладів для незалежності від мови написання документу та бази синонімів;

– *розбиття документу* на частини, визначення і підсумовування результату отриманого для пар частин як незалежних документів;

– *метод каскаду* – оптимізація може використовуватися для пошуку схожих документів серед проіндексованих даних. Полягає в кластеризації груп документів і виділення центрового документа. Негативний результат порівняння документу з центром кластера виключає потребу виконувати операції з документами що входять в кластер.

Основним недоліком будь-якого алгоритму знаходження подібності є його цільове призначення. Усі алгоритми суто орієнтовані на текст як суцільній елемент структурних даних і не враховують контексту інформації, яка представлена в тексті. Це унеможливлює застосування алгоритмів для тексту зі специфічним контекстом. Єдине застосування таких алгоритмів – це тексти, які можуть бути елементами більш-складного об'єкту порівняння.

У бібліографічному описі немає великих текстових елементів, отже застосування складних алгоритмів побудованих на шинглах чи супер-шинглах є недоцільним. Для бібліографічних описів найбільш практичним є застосування алгоритмів побудованих на метриках. З розглянутих таких алгоритмів найбільш ефективним є алгоритм визначення відстані редагування Дамерау-Левенштейна. Крім того, алгоритм можна зробити універсальним для будь-яких даних використовуючи підхід рефлексивно-орієнтованого програмування.

Задача ідентифікації подібності бібліографічних описів є похідною від задачі класифікації текстів, яка формулюється наступним чином: нехай є деяка множина прикладів текстів, кожен з яких належить до одного з k заздалегідь відомих класів. Потрібно створити алгоритм, який, будучи навченим на текстах-прикладах, отримуючи на вході новий невідомий текст, видавав на виході вектор $(p_1, ..., p_k)$, де p_i – ймовірність того, що даний текст належить класу і.

Узагальнюючи, можна зробити наступне формулювання: для заданого бібліографічного опису знайти підмножину бібліографічних описів, критерій подібності яких менший за деяке граничне значення [7].

3. Ціль та задачі дослідження

Проведені дослідження ставили за мету розроблення алгоритму пошуку подібності складно-структурованих даних, який забезпечує покращення якості ідентифікації подібності у вже існуючих програмних продуктах.

Для досягнення поставленої мети вирішувалися такі задачі:

– аналіз і побудова семантичної мережі бібліографічного опису публікації;

– розробка методів порівняння окремих вузлів семантичної мережі;

– розробка програмного забезпечення, яке реалізує розроблені методи порівняння та побудовану семантичну мережу для пошуку подібності складно-структурованих даних.

4. Матеріали та методи дослідження використання розробленого алгоритму для задачі пошуку подібності бібліографічних описів

В якості складно-структурованих даних обрано бібліографічний опис. Експериментальні дослідження ефективності розробленого підходу проведено для задачі пошуку подібності бібліографічних описів у системі звітності про наукову-дослідну діяльність Національного університету «Львівська політехніка» – інформаційно-аналітичній системі "ScienceLP" [8, 9].

Бібліографічний опис – це сукупність бібліографічних відомостей про документ, його складову частину чи групу документів, які наведені за певними правилами, необхідні та достатні, і є результатом аналітичного опрацювання інформації. Процес складання бібліографічного опису передбачає виявлення та формування за певною методикою множини бібліографічних даних про окремий документ або його частину чи групу документів.

Для загального складання бібліографічного опису на міжнародному рівні, використовується стандарт ДСТУ ГОСТ 7.1:2006 «*Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання*», який набув чинності 1 липня 2007 року. Він є

базовим для системи стандартів, правил, методичних посібників зі складання бібліографічного опису. Дані для складання описів беруться безпосередньо з видання. Опис складається з обов'язкових елементів: основний заголовок, автори, повторність видання, рік видання, обсяг [10].

Практичними застосуваннями методу пошуку подібності бібліографічних описів в системі звітності підрозділів є такі:

– пошук публікацій в базі даних за її бібліографічним описом;

– пошук подібності бібліографічних описів публікацій в базі даних.

5. Семантична мережа для бібліографічного опису

Семантична мережа – це спрямований граф з поіменованими вершинами і дугами, причому вузли позначають конкретні об'єкти, а дуги – відносини між ними. Семантичну мережу можна побудувати для будь-якої предметної області і для самих різноманітних об'єктів і відносин. Прикладом використання семантичних мереж для бібліографічного опису є робота [11], в якій розроблено систему для P2P обміну бібліографічними даними між науковцями.

Оскільки, бібліографічний опис можна подати у вигляді структури даних, де кожна його компонента є окремо виділена, можна значно підвищити якість пошуку подібності, якщо робити спеціалізований аналіз кожної його компоненти.

Для реалізації алгоритму порівняння складно-структурованих даних потрібно обробити дані таким чином, щоб можна було ідентифікувати кожен елемент, та яку функцію порівняння застосувати для нього. Для вирішення такої задачі найкраще підходить представлення бібліографічного опису у вигляді семантичної мережі, де кожна компонента бібліографічного опису є вузлом, і в залежності від складності цього компоненту цей вузол може мати дочірні вузли, які в свою чергу будуть теж ділитися на дочірні, поки весь опис об'єкту в семантичній мережі не буде представлений вузлами примітивних типів.

До кожного такого вузла буде застосовуватися функція порівняння. Також кожен вузол мережі має свій ваговий коефіцієнт, який розподіляється рівномірно між усіма меронімами одного холоніма та обчислюється як відношення коефіцієнта холоніма до кількості меронімів.

Для кожного вузла семантичної мережі в залежності від типу даних буде застосовуватися своя функція порівняння, яка на вхід буде отримувати два вузла однакового типу, а повертати їхнє значення подібності. Остаточне значення подібності для вузла визначається ваговим коефіцієнтом, на який і множиться значення подібності. Після чого усі значення вузлів поточного рівня вкладеності сумуються і отримане значення представляє подібність батьківського вузла. Отримане значення не може перевищувати значення вагового коефіцієнта.

Для різних примітивних типів можуть застосовуватися різні функції порівняння.

Числові типи даних. Для числових типів даних функція порівняння обчислюватиме відношення меншого числа до більшого. Таке відношення дасть представлення яку частину більшого числа представляє собою менше.

Стрічкові типи даних. Стрічка представляє собою набір символів у вигляді окремих слів. У даному випадку у бібліографічному описі стрічкові дані займають невеликий об'єм. Отже для ефективного порівняння доцільним є знаходження відстані редагування. На даний момент, найбільш ефективним алгоритмом знаходження відстані редагування є алгоритм Дамерау-Левенштейна. Ефективність полягає у великій кількості підтримуваних операцій в стрічці: вставка, видалення, заміна та перестановка.

Тип даних «дата». Для цього типу даних застосовується подібний механізм як до числових типів даних, лише з однією важливою відмінністю. Основна проблема при порівнянні дат – це часові межі. Для даного типу часові межі повинні бути відомими наперед. Для цього обчислюється найбільша різниця дат.

Таким чином можна дійти до висновку, що порівняння різних компонент залежить напряму від типу кожної компоненти. Використовуючи таку специфіку даних, реалізацію алгоритму можна зробити універсальною для будь-яких порівнюваних об'єктів.

Універсальність обчислювання можна досягнути за допомогою рефлексії. Під цим поняттям мається на увазі, що побудова семантичної мережі буде відбуватися саме на основі елементів структури та їхніх типів, порівнюваних об'єктів.

Множини. Окремими випадком є нечітке порівняння множин будь-яких структур даних з довільним порядком. Основна проблема полягає в обчислювальній складності такого алгоритму, оскільки передбачає декілька різних послідовних кроків, що виключає можливість використання динамічного програмування.

Нехай є дві множини: $Q\{a, ab, abc\}$ та $W\{b, abc, ba\}$. Першим кроком є обчислення вектору подібності для кожного елементу з множини Q. Тобто, елемент з множини Q порівнюється з кожним елементом з множини W і записуються у вектор. Після чого цей вектор відсортовується у порядку спадання знайдених подібностей елементів.

Далі для елемента множини Q обирається перший елемент з відповідного йому вектора, і проводиться пошук такого ж елемента серед перших елементів усіх векторів, причому подібність цих елементів може бути різною. Далі відбувається наступне:

– Якщо таких елементів не знайдено, то результат подібності встановлюється для поточного елемента множини Q і видаляється весь векто, з якого було обрано результат, та усі входження знайденого елемента у інших векторах.

– Якщо знайдено такі елементи, це означає, що елемент, який був обраний першим у векторі, може мати конфліктні подібності для інший елементів множини Q. Тому серед знайдених елементів серед перших елементів векторів спочатку шукається максимальне значення подібності. Якщо таке значення знайдено, то результат остаточно записується та видаляється весь вектор знайденого результату та усі входження знайденого елемента у інших векторах. Якщо ж знайдено декілька таких максимальних значень подібності серед однакових елементів вибірки з перших елементів векторів, то записується поточний результат вибраного

елементу і видаляється вектор, але усі інші елементи залишаються у векторах.

Далі знайдені результати подібностей множаться на внутрішній ваговий коефіцієнт колекції, який дорівнює відношенню одиниці до кількості елементів в множині, та сумуються.

Для розробки алгоритму порівняння складно-структурованих даних спочатку потрібно привести вхідні дані до універсального вигляду для подальшої їхньої обробки алгоритмом. Для цього було вирішено будувати семантичну мережу на основі структурних елементів порівнюваного об'єкту, використовуючи рефлексивно-орієнтований підхід.

6. Проектування семантичної мережі

Кожен вузол мережі представляє собою значення примітивного типу. Для кожного такого типу була розроблена власна функція порівняння. Для числових типів порівняння відбуватиметься на основі відношення меншого до більшого числа. Для стрічкових даних використовуватиметься алгоритм Дамерау-Левенштейна для обчислення відстані редагування. Для дат застосовуватиметься підхід на основі мінімальної та максимальної дати.

На рис. 1 подано побудовану семантичну мережу бібліографічного опису: вказано назву поля, тип даних та значення вагового коефіцієнта (ВК).

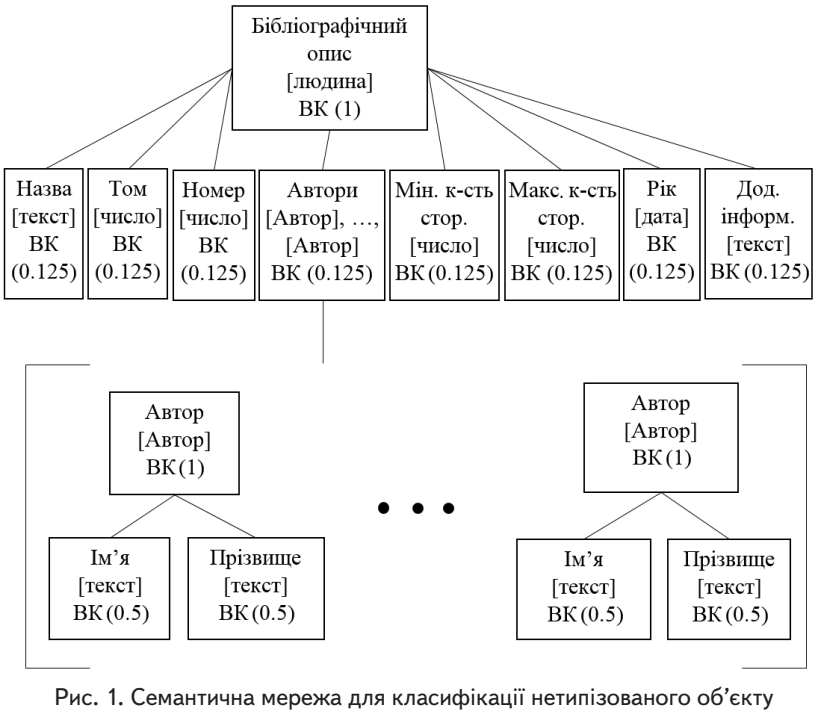

Рис. 1. Семантична мережа для класифікації нетипізованого об'єкту

У семантичній мережі визначення подібності бібліографічних описів кожна вершина найнижчого рівня представляє певний компонент бібліографічного опису, кожна вершина вищих рівнів представляє функцію порівняння, на вхід якої подаються відповідні частини двох бібліографічних описів, а на виході отримується коефіцієнт їх подібності. Кожна така функція має динамічний ваговий коефіцієнт, який визначається підсистемою під час порівняння бібліографічних описів і залежить від проміжних результатів. Елементи бібліографічного опису можуть одночасно подаватись на вхід різних функцій порівняння, наприклад, роки будуть порівнюватись на різницю і на поцифрову рівність. Коренева вершина видає результат подібності двох бібліографічних описів.

Аналіз подібності відбувається окремо за кожною компонентою бібліографічного опису: назва статті, рік видання, перелік авторів, місто, видавництво та кількість сторінок.

Оскільки рік та кількість сторінок є числовими даними, їх аналіз відбувається у двох напрямках: різниця чисел та поцифрове порівняння.

Список авторів аналізується також у двох напрямках: відбувається порівняння кількості авторів, співставляються імена кожного з авторів.

У семантичній мережі визначення подібності бібліографічних описів кожна вершина найнижчого рівня представляє певний компонент бібліографічного опису. Кожен вузол має свій ваговий коефіцієнт, який обчислюється в залежності від кількості вузлів на кожному з рівнів мережі.

7. Результати експериментальних досліджень пошуку подібності бібліографічних описів

На основі розробленої семантичної мережі та методів порівняння окремих її вузлів створено програмну реалізацію для дослідження процесу обробки інформації. Тестування програмного забезпечення проводилося на складно-структурованих даних типу «Бібліографічний опис». Об'єкт такого типу налічує у собі усі можливі структурні елементи, які уміє розпізнавати семантична мережа. Програмний модуль пошуку подібності бібліографічних описів публікацій впроваджено у інформаційно-аналітичну систему "ScienceLP" Національного університету «Львівська політехніка» [8]. Оскільки система "ScienceLP" містить персональні дані працівників університету, із роботою даного пошукового модуля можна ознайомитись із внутрішньої мережі університету, виконавши підключення до приватної мережі.

На рис. 2 наведено інтерфейс користувача ІАС "ScienceLP", на якому відображено приклад результатів пошуку подібності бібліографічних описів публікацій. Пошуковий запит відображається першим у результуючій вибірці даних. Для кожної публікації відображається її тип та повний бібліографічний опис. Результуюча вибірка даних відсортована у порядку спадання знайдених коефіцієнтів подібностей бібліографічних описів.

Для дослідження ефективності реалізованого алгоритму здійснено оцінку подібності для оригінального бібліографічного опису публікації та бібліографічних описів, які містять внесені зміни.

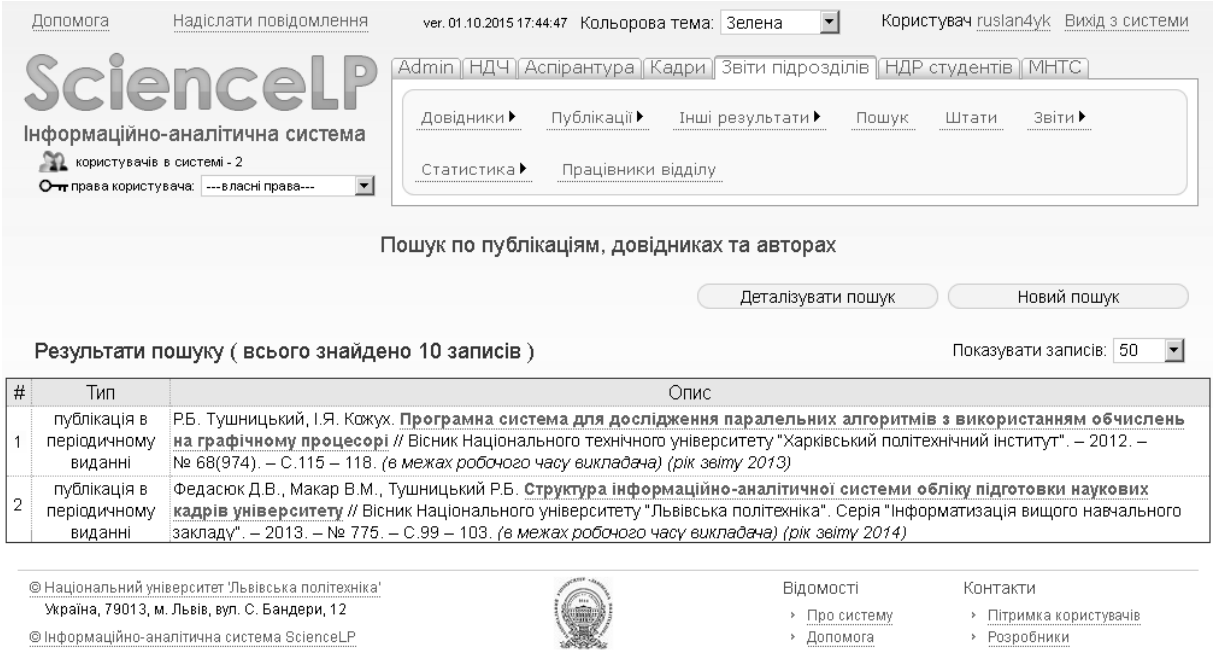

Рис. 2. Відображення результатів пошуку подібності бібліографічних описів публікацій в ІАС "ScienceLP"

В табл. 1 подано обчислений коефіцієнт подібності бібліографічних описів для такого оригінального опису:

Зайченко О. С., Петрушка І. М. "Особливості формування інформаційного забезпечення туристичної діяльності" / О. С. Зайченко, І. М. Петрушка // Вісник Національного університету "Львівська політехніка". Серія "Інформаційні системи та мережі". – № 783. – 2014. – с. 336-345.

Таблиця 1

Коефіцієнт подібності бібліографічних описів публікацій

Опис зміни	Змінений текст бібліографічного опису	Коефіцієнт подібності, %
Змінено порядок одного слова в назві	Особливості інформаційного формування забезпечення туристичної діяльності	90
Змінено порядок двох слів	Особливості діяльності інформаційного формування забезпечення туристичної	88
Внесено помилки в авторах	Зайчанко Ф. С., Петрушка К. М.	86
Внесено зміни в назві	Особливості аналізу програмного забезпечення туристичної діяльності	86
Внесено зміни у рік, номер та сторінки	№ 222. – 2015. – с. 111–222	67

З табл. 1 видно, що розроблений підхід пошуку подібності є стійким до внесення суттєвих змін у текстові значення бібліографічного опису і чутливим до внесення суттєвих змін у числові значення.

Великою перевагою універсальності розробленого рішення є застосування методу рефлексії. Такий підхід дає змогу розпізнати практично будь-який об'єкт, незалежно чи це вбудований, чи користувацький тип даних. Дослідження роботи цього підходу проводилося на модифікаціях типу «Бібліографічний опис», під час якого додавалися нові елементи, модифікувалися та видалялися існуючі. Це дозволило побачити процес побудови семантичної мережі в залежності від структури об'єкту. Було виявлено, що в залежності від кількості елементів, які характеризують об'єкт, по різному розподіляються коефіцієнти ваги вузлів. Значення такого коефіцієнту обчислюється пропорційно від кількості елементів об'єкта. Але можлива ситуація, коли коефіцієнт може представляти значення в періоді, для прикладу 0.333. Це може дати не суттєву похибку точності при виявленні подібності. На практиці виявлено, що розмір цієї похибки не впливає на кінцевий результат.

8. Обговорення результатів експериментальних досліджень

Порівняння кожного вузла семантичної відбувалося за заданими правилами, які встановлюються для кожного примітивного типу даних окремо. Завдяки тому, що кожен складний тип даних представляється у семантичній мережі як набір примітивних типів, кількість можливих правил є скінченною. Це дає змогу повністю перевірити порівняння кожного елемента об'єкту.

Дослідження швидкості та достовірності проводилося на вибірці даних у 1900 елементів.

Для числових типів серед усієї вибірки не було виявлено некоректних результатів. Це очевидно, оскільки алгоритм знаходження подібності для такого типу даних є простим та примітивним і займає всього одну базову операцію.

Для такого типу даних як дата, у ході дослідження було виявлено, що подібність таких елементів сильно залежить від періоду дати, в якому повинно відбуватися порівняння. Оскільки процес порівняння включає у собі переведення дат у єдину числову характеристику

та порівняння вже цих характеристик як звичайних чисел, то потрібно зауважити, що кожне число порівнюються від його початку відліку. Для звичайних чисел це є справедливе правило. Але для інших числових даних, які представляють інші типи, можуть бути додаткові обмеження. При порівнянні дат важливо знати точку відліку порівняння. Для цього потрібно знати мінімальне та максимальне значення дати у вибірці порівняння. Відповідно ці значення і встановлюють ліміти порівняння для досягнення більшої достовірності порівняння. Якщо ж вибірка складається усього з одного елементу, то при порівнянні це правило не повинно враховуватися, оскільки порівнюваний елемент та базовий якраз і відповідають за мінімальне і максимальне значення дати. Але є і недолік при такому підході, оскільки потрібно знати мінімальне та максимальне значення дати, то необхідно проводити попередній пробіг по вибірці порівнюваних даних для їх пошуку. А це в свою чергу збільшує обчислювальну складність алгоритму, що веде до збільшення часу пошуку подібності.

Як тільки алгоритм отримує на вході елемент складного типу, який уже представлений у семантичній мережі у вигляді елементів примітивних типів, відбувається рекурсивний виклик функції порівняння, який застосовує уже відому для кожного типу даних власну функцію порівняння. Такий рекурсивний виклик є необмеженим в залежності від складності структури порівнюваного об'єкта. Відповідно чим більша вкладеність складних структур тим більше часу необхідно для проведення порівняння.

Для вузлів семантичної мережі, які представляють колекції даних, процес порівняння ускладнюється кількістю елементів порівнюваних колекцій та появою колізій у результаті великої кількості однакових елементів у колекції. Враховуючи, що порівняння колекцій проводиться з врахуванням нечіткої подібності елементів, то стає складно визначити найбільш достовірний подібний елемент колекції, маючи при цьому ще таких самих декілька елементів. Можна зробити висновок, що проблема таких колізій обумовлена саме процесом нечіткого пошуку подібності елементів колекції.

На тестовій вибірці не вдалося знайти фактичне підтвердження існування такої проблеми, оскільки імовірність виникнення такої ситуації є надзвичайно мала, але теоретичне підґрунтя існує у цієї проблеми і це може бути потенційним матеріалом дослідження для покращення роботи алгоритму порівняння колекцій.

Отже, можна зробити висновок, що розроблений підхід виявлення подібності складно-структурованих даних містить виражений показник новизни, оскільки дає змогу проводити пошук подібності для об'єктів невідомої структури та типу.

Також явною ознакою новизни є те, що ефективність та достовірність результатів подібності бібліографічних описів є набагато вищою, чим достовірність результатів, виявлених в існуючій системі порівняння IAC "ScienceLP" [8, 9].

Але варто також зазначити, що є декілька недоліків алгоритму, які можуть порушувати точність результатів та збільшувати час отримання цих результатів. Цими недоліками є точка відліку при порівняння дат та проблема колізій при отриманні результатів подібності двох колекцій.

9. Висновки

Проведені дослідження показали, що існує проблема під час порівняння інформації. Суть проблеми полягає в тому, що сучасний підхід до збереження інформації все менше стає сумісним із старими методами пошуку подібності, оскільки інтелектуальне порівняння інформації існуючих систем дає змогу лише оперувати примітивними типами даних. Тому виникла необхідність розробити метод порівняння даних, який був би незалежний від типу даних. Основна задача методу – базуючись на структурі порівнюваного об'єкту, порівнювати кожну його компоненту в залежності від того, якого типу ця компонента є.

В результаті досліджень проведено аналіз і побудовано семантичну мережу бібліографічного опису публікації. Для окремих вузлів семантичної мережі розроблено свої методи порівняння, які базуються на типі вузла.

Розроблено програмне забезпечення, яке реалізує розроблені методи порівняння та побудовану семантичну мережу для пошуку подібності складно-структурованих даних. Для забезпечення універсальності запропонованого методу використано рефлексивно-орієнтований підхід програмування. Це дає змогу алгоритму бути незалежним від типу порівнюваного об'єкту та його внутрішньої структури.

Проведені експериментальні дослідження показали доцільність використання запропонованого алгоритму ідентифікації подібності складно-структурованих даних для системи звітності про наукову-дослідну діяльність Національного університету «Львівська політехніка».

Подальші дослідження включають в себе вдосконалення існуючих та розроблення нових функцій порівняння вузлів семантичної мережі.

Література

1. Broder, A. Z. On the Resemblance and Containment of Documents [Text] / A. Z. Broder // Proceedings of Compression and Complexity of SEQUENCES 1997, 1997. – P. 21–29. doi: 10.1109/sequen.1997.666900
2. O'Hara ,T. Lexical Acquisition with WordNet and the Mikrokosmos Ontology [Text] / T. O'Hara, K. Mahesh, S. Nirenburg // Proceeding of the COLING/ACL Worskshop on Usage or WordNet in Natural Language Processing Systems, 1998. – P. 94–101.
3. Nguyen, T. Combination of Lexical and Structure-Based Similarity Measures to Match Ontologies Automatically [Text] / T. Nguyen, S. Conrad // Knowledge Discovery, Knowledge Engineering and Knowledge Management. Communications in Computer and Information Science. – 2013. – Vol. 415. – P. 101–112. doi: 10.1007/978-3-642-54105-6_7
4. Metzler, D. Similarity Measures for Short Segments of Text [Text] / D. Metzler, S. Dumais, C. Meek // Advances in Information Retrieval. Lecture Notes in Computer Science. – 2007. – Vol. 4425. – P. 16–27. doi: 10.1007/978-3-540-71496-5_5

5. Metzler, D. Similarity measures for tracking information flow [Text] / D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, J. Zobel // Proceedings of the 14th ACM international conference on Information and knowledge management – CIKM '05, 2005. – P. 517–524. doi: 10.1145/1099554.1099695
6. Buttler, D. A Short Survey of Document Structure Similarity Algorithms [Text] / D. Buttler // The 5th International Conference on Internet Computing, 2004.
7. Ідентифікація бібліографічних описів [Електронний ресурс]. – 2015. – Режим доступу: https://uk.wikipedia.org/wiki/Ідентифікація_подібності_бібліографічних_описів
8. Макар, В. Інформаційно-аналітична система для автоматизації підготовки наукових звітів підрозділів Львівської політехніки [Текст]: матер. 6-ї наук.-прак. конф. / В. Макар, Р. Тушницький // Інноваційні комп'ютерні технології у вищій школі. – Львів, 2014. – С. 177–182.
9. Федасюк Д. В. Структура інформаційно-аналітичної системи обліку підготовки наукових кадрів університету [Текст] / Д. В. Федасюк, В. М. Макар, Р. Б. Тушницький // Вісник Національного університету "Львівська політехніка". Серія "Інформатизація вищого навчального закладу". – 2013. – № 775. – С. 99–103.
10. Кушнаренко, Н. М. Наукова обробка документів [Текст] / Н. М. Кушнаренко, Б. К. Удалова; 4-те вид., перероб. і доп. – К. : Знання, 2006. – 334 с.
11. Haase, P. A Bibster – A Semantics-Based Bibliographic Peer-to-Peer System [Text] / P. Haase, B. Schnizler, J. Broekstra, M. Ehrig, F. van Harmelen, M. Menken et. al. // Semantic Web and Peer-to-Peer, 2006. – P. 349–363. doi: 10.1007/3-540-28347-1_19

Розроблено інформаційну технологію, що базується на знаннях, яка вирішує задачу автоматичної генерації тестових запитань з групуванням їх відповідно до ієрархії понять предметної області. В рамках розробленої технології створено інструментальний програмний засіб. Розроблена технологія дозволить збільшити кількість навчальних тестів, звільнивши час викладача від рутинної роботи на користь її творчої складової, при цьому підвищить якісний рівень освіти

Ключові слова: електронне навчання, дистанційне навчання, навчальний контент, онтології, бази знань

Разработана базируемая на знаниях информационная технология, которая решает задачу автоматической генерации тестовых вопросов с группированием их в соответствии с иерархией понятий предметной области. В рамках разработанной технологии создано инструментальное программное средство. Разработанная технология позволит увеличить количество учебных тестов, освободив время преподавателя от рутинной работы в пользу её творческой составляющей, при этом повысит качественный уровень образования

Ключевые слова: электронное обучение, дистанционное образование, контент, онтология, базы знаний

УДК 004.853

DOI: 10.15587/1729-4061.2015.51334

РОЗРОБКА ГЕНЕРАТОРА ТЕСТІВ ДЛЯ "MOODLE" НА БАЗІ ОНТОЛОГІЇ

С. В. Сирота
Кандидат технічних наук, доцент*
E-mail:sergiy.syrot@gmail.com
В. О. Ліскін
Аспірант*
E-mail: lis-580@rambler.ru
*Кафедра прикладної математики
Національний технічний університет України
«Київський політехнічний інститут»
пр. Перемоги, 37, м. Київ, Україна, 03056

1. Вступ

Застосування інформаційних технологій у системі освіти дозволяє удосконалювати навчальний процес шляхом впровадження нових методів і підходів не тільки в навчанні, а й в контролі знань.

Стрімкий розвиток елементної бази та інформаційних технологій ставить завдання безперервно вдосконалювати, і тримати навчальний контент «up to date». Викладач працює над своїми курсами, використовуючи Інтернет, і редагує матеріали в реальному часі. Завдяки цьому студенти мають можливість централізовано і оперативно отримувати оновлену інформацію.

На сьогоднішній день тестування є однією з найбільш широко використовуваних форм перевірки знань. Одним з найяскравіших прикладів є ЗНО для випускників середніх шкіл, обов'язкове для вступу у ВНЗ з 2008 року та ДПА в середній школі.

Актуальною задачею є підвищення якості контролю знань. Аналіз методики роботи з тестовими запитаннями показав, що у випадку невеликої кількості банку запитань тести доцільно використовувати лише для фінального контролю в режимі екзамену. Звідси випливають дві полярні задачі. З одного боку тестування має виконувати навчальну функцію і бути максимально незалежним від випадковості, а з другого – реально відображати картину знань.