УДК 681.3 : 004.912

ПЕРСПЕКТИВИ АВТОМАТИЗОВАНОГО АНАЛІЗУ НАУКОВИХ ДОСЛІДЖЕНЬ

Кашкарьов А. О.

Таврійський державний агротехнологічний університет (м. Мелітополь)

Запропоновано спосіб виявлення відмінностей між науковими дослідженнями на основі автоматизованого аналізу тематичних структурованих текстів.

Постановка проблеми. Аналіз результатів наукових досліджень та робіт свідчить про беззаперечну ефективність використання електротехнологій у виробництві та переробці продукції сільського господарства, а також про наявну розрізненість наукових досліджень. Це пов'язано із специфікою впливу електротехнологій на біологічні об'єкти, широким переліком технологічних процесів (ТП), специфікою функціонування об'єктів керування та власними уподобаннями дослідників і виробничників. Автора особливо цікавить стан даного напряму використання елекротехнологій в умовах теплиць, що потребує ґрунтовного аналізу, для визначення ефективних технікотехнологічних рішень для впровадження та обґрунтування напрямів подальших наукових досліджень.

Наявні дослідження у рамках окремих технологічних процесів та впливу електротехнологій на біологічні об'єкти і ресурси (повітря, ґрунту, живильні розчини та ін.) розосереджені і здійснюються у рамках наукової спільноти або виробництва виходячи із певних традицій і ресурсної бази. Здійснити аналітичну оцінку цих напрацювань людськими ресурсами не можливо. Отже, постає завдання часткової автоматизації систематизації та аналізу інформації.

Аналіз останніх досліджень. Відомі методи аналізу наукової інформації ґрунтуються на людському факторі, що обумовлено специфікою подачі наукової інформації та результатів досліджень. Це не дозволяє виконати аналіз великих об'ємів інформації, з-за витрати часу та індивідуальних рис науковця. Автоматизований аналіз текстової інформації отримав поширення в автоматизованій семантичній оцінці текстів у сфері копірайтингу, для аналізу якості статей для електронних сторінок – SEO-аналіз текстів [8].

Відомі також методи формування класів, об'єктів, їх властивостей та методів в об'єктна-орієнтованому програмуванні (ООП) на основі опису принципу дії або технічного завдання [2, 3].

Раніше автором були запропоновані рекомендації для синтезу схем технологічних ліній малогабаритних комбікормових установок та побудови алгоритму керування ними на основі мереж Петрі [3], що дозволило класифікувати технологічне обладнання, синтезувати технологічні лінії та системи автоматичного керування ними. Вхідною інформацією для синтезу був рецепт комбікорму та умови підприємства.

Постановка завдання. Визначити інструментарій автоматизованого аналізу наукових досліджень та методи їх впровадження у виробництво.

Основні матеріали дослідження. Вирішити поставлене завдання можливо за допомогою автоматизованого аналізу результатів досліджень - автореферат, патент на корисну модель. Зазначені документи добре структуровані і формалізовані.

Вважаємо, що реалізувати поставлене завдання можливо за допомогою використання генетичних алгоритмів. Інструменти ООП дозволять автоматизувати зв'язок між елементами генетичного алгоритму (ГА) (визначення хромосом, генерація особин і популяцій, визначення пристосованості та ін.) та сформувати бази вхідних даних.

Враховуючи їх чисельність та різноманітність , а також тенденції до міждисциплінарних наукових досліджень і результатів, то необхідно відзначити, що розв'язати поставлене завдання можливо тільки за рахунок автоматизованого аналізу.

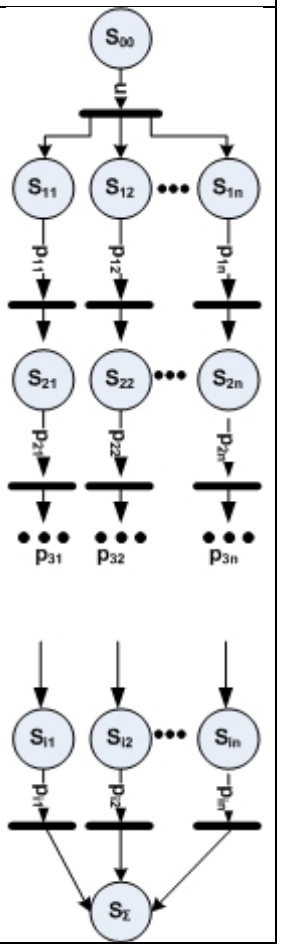
Пропонуємо аналізувати структуровані тексти, які знаходяться у вільному доступі: патенти на корисні моделі (Український інститут інтелектуальної власності), автореферати дисертацій (Національна бібліотека України ім. Вернадського).

Як було зазначено вище, компанії замовники статей, компанії та програми, які індексують електроні сторінки давно використовують SEO-аналіз [8]. Він дозволяє утримувати базу даних ключових слів, архівувати інформацію та здійснювати швидкий доступ до неї. Наукове використання інструментарію полягає у маркетингових та соціологічних дослідженнях. Але дієві алгоритми та результати роботи здебільшого представляють технічну та економічну таємницю. Тому, для аналізу відкритих структурованих наукових текстів та практичного використання результатів досліджень пропонується використовувати симбіоз:

- бази даних та система керування ними;
- частотний та SEO-аналіз текстів;
- об'єктно-орієнтоване програмування;
- функціональні мережі Петрі (ФМП);
- генетичні алгоритми.

Найбільш розповсюджені структуровані документи, які містять наукову та науково-виробничу інформацію: автореферат, патент на корисну модель. Зазначені документи можливо представити як об'єкт мережі Петрі [7] у вигляді графів (табл. 1, табл.2), що дозволить використати функціональні мережі Петрі. На відміну від відомих методів дозволить вирішити мінімаксні задачі та задачі оптимізації, з метою подолання технічних протиріч та неточностей [1]. Зазначене можливо за допомогою використання функціональних мереж Петрі (МП), ГА, формування баз даних результатів наукових досліджень та їх аналізу за проблемно-орієнтованим напрямом.

Таблиця 1 - Структура автореферату

i	Найменування розділу (властивість об'єкту)	Відображення у вигляді елементів стохастичної / функціональної МП
1	Назва роботи (мета запиту)	
2	Актуальність	
3	Зв'язок роботи з науковими програмами	
4	Мета роботи	
5	Задачі досліджень	
6	Об'єкт досліджень	
7	Предмет дослідження	
8	Методи дослідження	
9	Наукова новизна	
10	Практичне значення	
11	Особистий внесок	
12	Апробація	
13	Основний зміст роботи	
14	Висновки	
15	Анотація	
16	Список опублікованих праць	

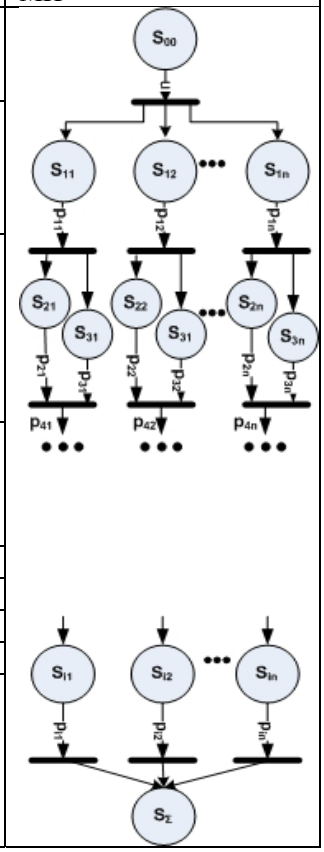
Примітка: n – кількість документів для аналізу

Для здійснення досліджень за представленим напрямом необхідно сформувати систему управління базою даних структурованих текстів (БДСТ), яка міститиме інформацію про авторів, інформаційні реквізити документа та заповнені структурні елементи (табл. 1, табл. 2). Записи БДСТ, характеризуються як класи, з якого формуються об'єкти із спільними властивостями та методами. Властивості об'єкту повинні відповідати структурі документі (табл. 1, табл. 2).

Методи об'єкту повинні здійснювати аналіз окремого запису БДСТ, властивості об'єкту, а також групи записів БДСТ. На даному етапі досліджень пропонується, що вказані методи складатимуться із функцій SEO-аналізу текстів:

- статистика тексту (кількість знаків, слів, частота слів, таблиці, рисунки та креслення);
- семантичне ядро, ключові слова та їх щільність;
- "водяність" (стоп-слова), "нудотність" та "академічна нудотність" тексту;
- заспамленість тексту;
- унікальність тексту та унікальні фрази.
- змішані слова та слова із різною розкладкою на клавіатурі.

Таблиця 2 – Структура документу з охорони інтелектуальної власності (патент на корисну модель)

i	Найменування розділу (властивість об'єкту)	Відображення у вигляді елементів стохастичної МП
1	Назва патенту (мета запиту)	
2	Галузь застосування	
3 4	Реферат: - частина спільна - частина відмінностей	
5 6	Аналог / прототип - опис - недоліки	
7	Задача	
8	Реалізація задачі	
9	Креслення	
10	Опис у статиці	
11	Опис у динаміці	

Примітка: n – кількість документів для аналізу

Аналіз елементів класу та відповідних записів які складають БДСТ, здійснюється на основі використання елементів статистики та граматичних правил. У свою чергу, аналіз БДСТ здійснюється на основі математичного апарату множин. Таким чином, зазначена концепція відповідає основним парадигмам ООП (інкапсуляція, успадкування, поліморфізм та абстрагування), що дозволяє використання об'єктного моделювання на основі МП [2, 7].

SEO-аналіз записів БДСТ визначить спільні та відмінні роботи за гістограмою ключових слів [8]. Цей етап також наповнить хромосоми пошуку та оцінки технічних рішень апаратом генетичних алгоритмів. Декомпозиція записів БДСТ за їх властивостями оптимізує пошук оптимальних технічних та технологічних рішень вирішення технічних протиріч [1].

Оптимізація обчислень, підвищення швидкості обробки даних, візуалізація результатів запиту та динамічне його корегування можливе за допомогою математичного апарату МП [6]:

$$N=(P, T, F, W, M_0) \quad (1)$$

де P – не порожня множина елементів мережі - вузли;

T – не порожня множина елементів мережі, які називають переходами;

$F \subseteq P \times T \cup T \times P$ – відношення інцидентності, ініціалізація властивостей та доступу до них, для *(P, T, F)*;

$W: F \to N \setminus \{0\}$ – кратність дуг.

$M_0: P \to N$ – початкова розмітка. Кожному вузлу $p \in P$ призначається деяке число $M_0(p) \in N$.

Термінологія МП матиме такі доповнення, які відповідатимуть аналізу записів БДСТ:

- P – не порожня множина елементів властивостей записів у БДСТ;
- T – не порожня множина точок запиту результатів реалізації методів об'єктів класу або інші функціональні умови переходу;
- $F \subseteq P \times T \cup T \times P$ – взаємодія результатів аналізу множини P та подальших дій за допомогою T;
- $W: F \to N \setminus \{0\}$ – приймаємо кратність всіх дуг $n=1$ – ординарна мережа.
- $M_0: P \to N$ – початкова розмітка характеризує заповнення властивостей об'єктів (записи БДСТ), або характеризує результати аналізу попередніх запитів за відповідними методами.

Математичний апарат МП та архітектура БДСТ на її основі дозволить проводити складний автоматизований семантичний аналіз групи текстів, не втрачаючи їх батьківську інформацію кожного об'єкту. Якщо говорити термінологію генетичних алгоритмів, то МП дозволять визначити документи (об'єкти), які відповідають мутаціям батьківських популяцій (технічне завдання або пошуковий запит), що направлені на вирішення наукових та практичних проблем.

Для програмної реалізації пропонованого підходу доцільно використовувати математичний апарат функціональних і кольорових сіток Петрі, що є достатньо тривіальною задачею. Більш перспективним з наукової та практичної точки зору є використання мереж Слєпцова [4] (розширення МП), що може сприяти прискоренню обчислень [5] реалізації паралельного SEO-аналізу властивостей об'єктів (записів БДСТ), корегування умов пошуку та ранжування за релевантністю.

Обробка представленої інформації повинна забезпечувати частотний та SEO-аналіз, як елемента структури, так і документа в цілому. Слід акцентувати увагу на концепції ООП, як інструменту додаткових даних щодо ініціалізації початкової популяції у генетичному алгоритмі.

Висновок. Для вирішення поставленого завдання запропоновано симбіоз використання баз даних, SEO-аналізу текстів, об'єктна-орієнтованого програмування та представлення структурованих наукових текстів, функціональних мереж Петрі та генетичних алгоритмів.

Такий підхід надасть можливість використовувати різні алгоритми оцінювання особин, формування батьківської популяції, рекомбінації, моніторингу мутацій із збереженням батьківської інформації кожного запису у базі даних структурованих текстів.

Відкриті джерела авторефератів дисертаційних досліджень та патентів на корисні моделі дозволяють сформувати базу даних тематичних батьківських популяцій наукових та виробничих рішень, що може стати стартом програми автоматичного подолання технічних протиріч та генерації винаходів.

Список використаних джерел

1. Альтшуллер Г. С. Поиск новых идей: От озарения к технологии: (Теория и практика решения изобретательских задач) / Г. С. Альтшуллер., Б. Л. Злотин, А. В. Зусман, В. И. Филатов. — Кишинев: Картя молдавеняскэ, 1989. — 382 с.
2. Бобровский С. И. Delphi 7. Учебный курс / С. И. Бобровский – СПб.: Питер, 2005. – 736 с.
3. Діордієв В. Т. Синтез та формалізація технологічних схем приготування комбікормів на малогабаритних комбікормових установках / В. Т. Діордієв, А. О. Кашкарьов // Праці Таврійського державного агротехнічного університету. – Мелітополь: ТДАТУ, 2008. – Вип. 8, Том 5. – С. 26-36.
4. Зайцев Д. А. Вычисления на сетях Слепцова / Д. А. Зайцев // Системная информатика, 2017, № 9, С. 42-62: - Режим доступа: http://www.system-informatics.ru/ru/article/145
5. Зайцев Д. А. Парадигма вычислений на сетях Петри / Д. А. Зайцев // Автоматика и телемеханика, - № 8, 2014, с. 19-36. : - Режим доступа: http://mi.mathnet.ru/at14104
6. Котов В. Е. Сети Петри / В. Е. Котов. – М.: Наука, 1984. – 160 с.
7. Стеценко И. В. Теоретические основы Петри-объектного моделирования систем / И. В. Стеценко // Математичні машини і системи. – 2011, № 4. – С. 136-148. – Режим доступу: www.immsp.kiev.ua/publications
8. Шардаков Д. Практический копирайтинг и маркетинг: основы, секреты и примеры от Даниила Шардакова / Д. Шардаков [Электронный ресурс]: - Режим доступа: https://shard-copywriting.ru/guest/giga-post-seo-analiz-teksta-11-instrumentov-dlya-kopiraytera.

Аннотация

ПЕРСПЕКТИВЫ АВТОМАТИЗИРОВАННОГО АНАЛИЗА НАУЧНЫХ ИССЛЕДОВАНИЙ

Кашкарёв А. А.

Предложен способ определения отличий между научными исследованиями на основе автоматизированного анализа тематических структурированных научных текстов.

Abstract

PROSPECTS OF AUTOMATED ANALYSIS OF SCIENTIFIC RESEARCH

Kashkarov A.

Proposed a method for determining the differences between scientific research on the basis of automated analysis of thematic structured scientific text.