Міністерство освіти і науки України
Національна металургійна академія України

КУРОП'ЯТНИК ОЛЕНА СЕРГІЇВНА

УДК [510.8+519.17]:004.91

КОНСТРУКТИВНО-ПРОДУКЦІЙНІ МОДЕЛІ ПРИРОДОМОВНИХ ТЕКСТІВ ДЛЯ ВИЯВЛЕННЯ ЗАПОЗИЧЕНЬ У СТРУКТУРОВАНИХ ДОКУМЕНТАХ

01.05.02 – математичне моделювання та обчислювальні методи

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата технічних наук

Дніпро – 2020

Дисертацією є рукопис.

Робота виконана на кафедрі комп'ютерних інформаційних технологій Дніпровського національного університету залізничного транспорту ім. академіка В. Лазаряна (м. Дніпро) Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Шинкаренко Віктор Іванович,
завідувач кафедри комп'ютерних інформаційних технологій Дніпровського національного університету залізничного транспорту ім. академіка В. Лазаряна, м. Дніпро

Офіційні опоненти: доктор технічних наук, професор
Шаронова Наталія Валеріївна,
завідувач кафедри інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут», м. Харків

доктор технічних наук, професор
Кіріченко Людмила Олегівна,
професор кафедри прикладної математики Харківського національного університету радіоелектроніки, м. Харків

Захист відбудеться «15» квітня 2020 р. о 12.00 годині на засіданні спеціалізованої вченої ради Д 08.084.01 у Національній металургійній академії України за адресою: 49600, м. Дніпро, пр. Гагаріна 4, НМетАУ, конференц-зал засідань.

З дисертацією можна ознайомитись у бібліотеці Національної металургійної академії України за адресою: 49600, м. Дніпро, пр. Гагаріна, 4, НМетАУ.

Автореферат розіслано « 11» березня 2020 р.

Вчений секретар
спеціалізованої вченої ради Д 08.084.01
к.т.н., доцент Т. В. Селівьорстова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність роботи. Однією з задач обробки текстів є їх зіставлення з метою виявлення запозичень. Її вирішення ускладнюється стрімким та перманентним зростання обсягів інформації у вільному доступі. Окремо слід відзначити ускладнення, пов'язані з маскуванням – спробами приховати запозичення, використовуючи технічні прийоми (зміни регістру, кодування, порядку абзаців тощо).

В останні роки в академічному середовищі все гостріше постає проблема неправомірних запозичень в навчальних та кваліфікаційних роботах – академічного плагіату, що призводить до зниження рівня якості вищої освіти та підготовки фахівців.

Однією з особливостей виявлення запозичень у кваліфікаційних роботах є необхідність врахування структури документів – наявності упорядкованих розділів та підрозділів, що пов'язані за змістом та мають особливості форматування. Дана структура впливає на добір матеріалів для зіставлення та інтерпретацію результату, оскільки кожен з розділів може мати свій допустимий відсоток запозичень через можливість вмісту опису відомих методів, алгоритмів тощо.

На сьогодні Законом України «Про вищу освіту» визначено, що «Виявлення … академічного плагіату є підставою для відмови у присудженні відповідного наукового ступеня.» (ст. 6, п. 6), а також «Система забезпечення закладами вищої освіти якості освітньої діяльності та якості вищої освіти (система внутрішнього забезпечення якості) передбачає здійснення таких процедур і заходів:… забезпечення функціонування ефективної системи запобігання та виявлення академічного плагіату» (ст. 16). Відповідно ст. 32, п. 3 «Заклади вищої освіти зобов'язані: 1) вживати заходів, у тому числі шляхом запровадження відповідних новітніх технологій, щодо запобігання та виявлення академічного плагіату в наукових роботах наукових, науково-педагогічних, педагогічних, інших працівників і здобувачів вищої освіти…».

Таким чином, розв'язання задачі виявлення плагіату, а в більш широкому значенні та неправовому полі – запозичень, є затребуваним на законодавчому рівні України.

Часткове вирішення задачі виявлення запозичень можливе за допомогою впровадження відповідних інформаційно-технічних та технологічних засобів.

Існує велика кількість технічних засобів: web-сервісів та Windows-додатків для виявлення запозичень. Вони різняться наявністю/відсутністю орієнтації на академічне середовище та підтримкою мов документів, що перевіряються.

Розробка відповідних комп'ютерних програм, які прийнято називати антплагіатами, потребує в першу чергу розробки моделі об'єктів перевірки (цифрового представлення структурованих документів) та методу порівняння їх контенту (тексту).

Моделям представлення текстів та мов присвячено велику кількість робіт в області штучного інтелекту, в тому числі text mining, комп'ютерної лінгвістики та natural language processing (NLP), виявлення плагіату (plagiarism detection, PD). Науковці світу, такі як S. Clark, P. Rychlý, Y. Bengio, P. F. Brown та ін. займаються проблемами представлення природної мови у формалізованому вигляді. Серед вітчизняних робіт з особливостей використання та сприйняття природної мови та NLP слід

відзначити наукові доробки Ю. В. Крака, О. В. Бісікала, О. О. Марченка, Н. В. Шаронової, А. В Анісімова, А. В Широкова, Я. О. Кохана, Ю. Р. Валькмана та ін.

В області NLP, пов'язаної з PD, розроблено відповідні моделі, методи та алгоритми, описані в роботах A. H. Osman, Л. А. Лупаренко, M. Mozgovoy, T. Kakkonen, E. Stamatatos, І. В. Груздо, І. В. Шостака, N. Salim. В області PD також розвивається напрям обробки інформації, представленої штучними мовами, в тому числі математичні обчислення та тексти програм, над яким працюють I. Smeureanu, M. Joy, G. Cosma, А.Н. Мироненко, Ш. Курмангалєєв, О. Н. Шиков та ін.

Існуючі моделі, методи та алгоритми застосовуються розрізнено і, переважно, не мають академічної спрямованості. Вони не враховують структурні особливості документів при доборі матеріалів та їх зіставлення. З огляду на великі обсяги документів, наявність у них фрагментів різними мовами доцільним є розробка моделей текстів, методів та засобів виявлення запозичень на їх основі з урахуванням структурних особливостей.

Зв'язок роботи з науковими програмами, темами, планами. Робота виконана відповідно до Закону України № 2623-14 від 11.07.2001 «Про пріоритетні напрями розвитку науки і техніки», Закону України «Про вищу освіту» (документ 1556-VII, редакція від 09.08.2019).

Дисертація є частиною науково-дослідних робіт, виконаних на кафедрі комп'ютерних інформаційних технологій Дніпровського національного університету залізничного транспорту ім. академіка В. Лазаряна, зокрема «Прикладне конструктивне моделювання програмних сутностей» (2016 р. № держреєстрації 0116U006841) та «Підвищення конкурентоспроможності залізничного транспорту на основі уніфікованих інтелектуальних технологій процесів перевезень та експлуатації парків технічних систем» (2017-2018 р. № держреєстрації 0117U004392), у яких автор брала участь як відповідальний виконавець і виконавець відповідно.

Мета та задачі дослідження. Метою дисертації є розробка моделей природомовних текстів, методу і засобів виявлення запозичень на їх основі у структурованих документах з маскуванням.

Для досягнення мети були поставлені та вирішені задачі дисертаційного дослідження. Необхідно було розробити:

- конструктивно-продукційну модель мови для формалізації семантичних аспектів порівняння тестів;
- конструктивно-продукційну та об'єктно-орієнтовану моделі мовних конструкцій (текстів) та метод їх порівняння з урахуванням структурних особливостей документів, яким вони належать;
- модель процесів маскування запозичень та її комп'ютерну реалізацію;
- алгоритмічне та програмне забезпечення для виявлення запозичень у структурованих документах.

Об'єктом дослідження є процеси виявлення та маскування запозичень у структурованих документах.

Предметом дослідження є конструктивно-продукційні моделі мови та текстів для виявлення запозичень у структурованих документах, методи та засоби їх реалізації, орієнтовані на використання в академічному середовищі.

Основою моделей є конструктивні властивості формальних граматик як деяких типів числень (за метаматичною енциклопедією) та інтерпретація операцій через алгоритми їх виконання.

Методи дослідження. Для вирішення поставлених задач було використано: конструктивно-продукційне моделювання, що базується на використанні апарату формальних мов та граматик і знайшло відображення у моделюванні за допомогою конструкторів та застосуванні методів їх перетворення; теорію графів (способи представлення графів, методи стиснення графів на основі SNAP-алгоритмів, операцій гомеоморфізму), методи та засоби теорії множин.

При розробці програмної реалізації побудованих моделей була застосована технологія об’єктно-орієнтованого проектування та програмування з використанням методів ефективної організації, пошуку та сортування інформації для складних структур даних, а також CASE-технології, в основі яких є уніфікована мова моделювання UML. Для дослідження часової ефективності програмної реалізації розроблених моделей використано статистичні методи, зокрема регресійний аналіз.

Наукова новизна отриманих результатів. В роботі **вперше отримано** такі результати:

1) виконано формалізацію процесів формування образів людини засобами об’єктно-орієнтованого моделювання та розроблено конструктивно-продукційну модель природної мови на основі образного представлення дійсності. Модель відрізняється від існуючих можливістю опису операцій модифікації мови і врахуванням внутрішнього виконавця, що може бути використано для продукування конструкцій різної складності та виявлення семантичних запозичень;
2) розроблено конструктивно-продукційну модель графового представлення текстів та метод їх порівняння для виявлення запозичення з урахування зміни порядку текстових складових та структурних особливостей документів, яким вони належать;
3) побудовано конструктивно-продукційну модель процесів маскування запозичень для автоматизації перевірки здатності демаскування запозичень у текстах програмами-антиплагіатами.

Отримали подальший розвиток:

4) методи та засоби конструктивно–продукційного моделювання: визначено зв'язок конструктивних моделей з об’єктно-орієнтованими, представленими засобами UML, який покладено в основу комп’ютерних реалізацій моделі графового представлення тексту для виявлення запозичень у структурованих документах та моделі процесів маскування;
5) методи виявлення семантичних запозичень у текстах: інтерпретація семантичної складової текстів виконавцем моделі природної мови дозволяє зменшити вплив зміни лексичної та синтаксичної структури тексту на виявлення запозичень;
6) методи обробки графів: на основі операцій гомеоморфізму та алгоритму групування вузлів за атрибутами розроблено метод стиснення графа з текстовим навантаженням.

Практичне значення отриманих результатів. В ході розробки конструктивно-продукційної моделі природної мови, основаної на образному сприйнятті світу людиною, було формалізовано процеси мислення, які нерозривно пов'язані з кодуванням і передачею думок за допомогою елементів спілкування – жестів, міміки, мови, писемності. Останні є основою для визначення вільної мови і індивідуальної мови людини. Модель дозволяє:

- розглядати природну мову як сукупність комунікативних здібностей людини;
- розглядати мову як конструктивний процес, що може бути покладено в основу створення методології побудови систем з високим ступенем інтелектуальності;
- формально представити класифікації мови (ареальну, по сфері вживання (загальновживаний, професійний)) з урахуванням особливостей її носія;
- удосконалювати семантичну NLP, зокрема в задачах зіставлення і виявлення запозичень у текстах, в значній мірі зменшуючи вплив синонімів, омонімів, перефразування, перекладу.

Можлива область застосування представленої моделі охоплює NLP-компоненти роботів і додатків, в тому числі систем перекладу і антиплагіату, експертних систем.

Розроблена модель процесів маскування дає формалізоване представлене відповідних процесів для подальшої автоматизації і дозволила отримати інструментарій, який підлягає налаштуванню, для отримання бази тестів для програм-антиплагіатів.

Розроблено алгоритмічні та програмні засоби виявлення запозичень у структурованих документах на основі моделі графового представлення тексту. Вони були використані для перевірки наявності запозичень у структурованих документах – дипломних проектах зі спеціальності 121 «Інженерія програмного забезпечення».

Практичне значення результатів підтверджується:

- актами впровадженням результатів дисертаційної роботи та прийняття у дослідну експлуатацію програмного засобу «Система виявлення запозичень у цифровому представлені структурованих документів» («StructuredDoc–Comparison») в Дніпровському національному університеті ім. академіка В. Лазаряна;
- актом впровадження результатів дисертаційної роботи в філії «Проектно-конструкторське бюро інформаційних технологій» АТ «Українська залізниця» при проектуванні та розробці алгоритмів та програмних засобів порівняння та видалення тотожних частин документів;
- актом впровадження результатів дисертаційної роботи в ТОВ «СОВЛАНУТ» при виконанні проекту «Розробка системи виявлення запозичень»;
- свідоцтвом про реєстрацію авторського права на твір № 68137 від 05.10.2016;
- свідоцтвом про реєстрацію авторського права на твір № 87131 від 22.03.2019;
- використанням отриманих результатів у НДР № держреєстрації 0116U006841 та 0117U004392.

Достовірність та обґрунтованість результатів та рекомендацій підтверджується зіставленням сучасних наукових і технічних досягнень в області моделювання мов і розробки алгоритмів та систем виявлення запозичень; базується

на коректному застосуванні відомих методів теорії графів, теорії множин, формальних граматик, статистичних методів, застосованих для вирішення коректно поставлених задач; апробацією основних теоретичних і експериментальних результатів роботи в друкованих працях та доповідях на конференціях та наукових семінарах.

Розроблені конструктивно-продукційні моделі покладені в основу програмних засобів для виявлення запозичень у структурованих документах та формування тестів для перевірки здатності демаскування запозичень. Виконано дослідження часової та функціональної ефективності алгоритмів зіставлення документів, отримані результати зіставлено з аналогом.

Особистий внесок здобувача. Основні результати дисертаційної роботи опубліковано в статтях у співавторстві:

[1] – розробка графової моделі тексту, розробка програмного забезпечення з її використанням та виконання її SR-оцінювання;

[2] – виконання інтерпретації відомих представлень смислу і семантики засобами уніфікованої мови моделювання, використання принципів об'єктно-орієнтованого проектування для формалізації мислення, думки та слова;

[3] – розробка конструктивно-продукційної моделі тексту та його графового представлення засобами конструктивно-продукційного моделювання, визначення правил стиснення для графа з текстовим навантаженням;

[4] – визначення основних проблем в задачах виявлення плагіату і використанні автоматизованих засобів для їх вирішення; аналіз і систематизація інформації, отриманої в ході огляду, тестування та аналізу роботи існуючих систем виявлення запозичень;

[5] – огляд та аналіз існуючих моделей мови, розробка конструктивної моделі мови, дослідження та аналіз придатності моделі для виявлення семантичного плагіату;

[6] – виконано формалізацію засобів перевірки здатності демаскування запозичень в програмах виявлення плагіату, розроблено конструктивно-продукційну модель процесів маскування у вигляді конструкторів сценаріїв модифікації текстів та процесу їх застосування.

В одноосібній статті [7] – модифікація конструктивної моделі графового представлення тексту, розробка об'єктно-орієнтованої (ОО) моделі тексту, визначення зв'язку між конструктивною та ОО-моделями, розробка алгоритмічного і програмного забезпечення для виявлення запозичень, дослідження часової ефективності комп'ютерної реалізації моделей.

Апробація результатів дисертації. Результати дисертаційної роботи представлено на всеукраїнських та міжнародних науково-практичних конференціях: «Сучасні інформаційні технології на транспорті, в промисловості та освіті» (Дніпро, 2012, 2013), «Інформаційні технології в металургії та машинобудуванні» (Дніпро, 2015, 2018, 2019), «Проблеми економіки транспорту» (Дніпро, 2015), «Проблеми математичного моделювання» (Кам'янське, 2015, 2018), «Комп'ютерне моделювання та оптимізація складних систем» (Дніпро, 2015), «Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті» (Дніпро, 2015 – 2018), «Теоретичні та прикладні аспекти побудови програмних систем» (Київ, 2016), «Інформаційні технології в моделюванні» (Миколаїв, 2017),

науковому семінарі кафедри комп'ютерних інформаційних технологій ДНУЗТ ім. акад. В. Лазаряна (м. Дніпро, 2013-2019), регіональному науковому семінарі «Проблеми розвитку інформаційних технологій і систем залізничного транспорту» (Дніпро, 05.06.18), регіональному науковому семінарі Придніпровського наукового центру НАН України «Сучасні проблеми керування та моделювання складних систем» (Дніпро, 25.09.19), науковому семінарі кафедри інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут» (Харків, 03.10.19).

Публікації. Результати дисертаційної роботи опубліковані в 24 наукових працях, в тому числі 7 статей у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях: «Проблеми програмування» (передрук у CEUR Workshop Proceedings – проіндексовано міжнародною науково-метричною базою (НМБ) **Scopus**), «Штучний інтелект», «Східно-Європейський журнал передових технологій» (НМБ **Index Copernicus**), «Наука та прогрес транспорту» (НМБ **Index Copernicus**), «Acta Cybernetica» (Угорщина, НМБ **Scopus, Web of Science**) , «Information Technologies & Knowledge» (Болгарія), «Системні технології» (НМБ **Index Copernicus**); у тезах доповідей та трудах міжнародних та всеукраїнських конференцій – 15.

Отримано 2 свідоцтва про реєстрацію авторського права на твір.

Структура та обсяг дисертації. Дисертаційна робота складається із вступу, чотирьох розділів, висновків, додатків і списку використаних джерел. Загальний обсяг дисертації становить 207 сторінок, в тому числі 161 сторінку основної текстової частини, 47 рисунків, 14 таблиць, 3 додатки на 12 сторінках та список використаних джерел із 128 найменувань на 16 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі на основі аналізу сучасного стану засобів формалізації мови та мовних конструкцій, а також методів та засобів виявлення запозичень обґрунтовано актуальність теми дисертаційної роботи, сформульовано мета та задачі, визначено об'єкт і предмет досліджень, наукову новизну й практичну цінність отриманих результатів, методи досліджень.

У першому розділі визначено проблеми виявлення запозичень у структурованих документах, виконано огляд та аналіз існуючих сучасних підходів до моделювання різних аспектів мови, методів, алгоритмів та програмних засобів (ПЗ) попередньої обробки та зіставлення текстів.

Проаналізовано такі підходи до моделювання як функційні мови, апарат керуючих просторів, n-грами, прагматичні типи, аналіз конструкцій на основі образів, моделювання на основі онтології та мереології. Дані підходи у різному ступені дозволяють охопити лексику мови, її семантику та синтаксис. До виявлених недоліків можна віднести повну або часткову відсутність можливості побудови принципово нових мовних конструкцій, а також наявність великої кількості параметрів, що ускладнює роботу з моделлю в умовах еволюції мови.

Проведений аналіз підтверджує, що розглянуті питання моделювання мови і її конструкцій є міждисциплінарними і наразі мають два напрями вирішення: гуманітарний та технічно-прикладний. Для першого характерним є широке охоплення,

проте низький ступінь формалізації, що ускладнює використання напрацювань для вирішення прикладних задач. Другий напрям відзначається виключно практичною спрямованістю, проте не завжди враховує особливості людини як носія мови.

Розглянуто методи та алгоритми обробки мовних конструкцій (текстів) для виявлення запозичень. Методи, засновані на простому порівнянні (відбитки, жадібне порівняння та ін.), є чутливими до механізмів маскування запозичень. Вивчення таких алгоритмів вказує на те, що актуальною є задача не лише зіставлення, а й попередньої оброки, необхідність якої зумовлена об'ємами інформації та можливістю використання технік маскування; а також аналізу отриманих результатів.

Розглянуті моделі, методи та алгоритми не враховують структурні особливості цифрового представлення документів, яким належить текст, що оброблюється. Серед розглянутих ПЗ не виявлено таких, що враховують структуру документів, текст яких перевіряється на унікальність.

На основі виконаного аналізу сформульовані задачі дисертаційного дослідження.

Другий розділ присвячено об'єктно-орієнтованому моделюванню смислових конструкцій та конструктивно-продукційному моделюванню природної мови, графового представлення тексту, методам стиснення графа з текстовим навантаженням та порівняння структурованих документів.

Для моделювання смислових конструкцій як результату розумової діяльності людини визначено такі поняття:

– прообраз – сутність, об'єкт, процес, подія, явище матеріального або віртуального (абстрактного) світу або окрема їх властивість. Це частина світу, що розглядається відокремлено від решти;
– прообраз особистого – позначення думок і емоцій людини, реакцій на те, що відбувається, а також деякі дії, які виражають ставлення до подій;
– смисл (інтенсіонал) – уявлення про прообраз. Носій смислу – почуття, дія, думка;
– семантика – це множина пов'язаних між собою тріад: прообраз – слово – смисл;
– сигніфікат – набір ознак, за якими можна чітко ідентифікувати предмет (явище), щоб його правильно назвати;
– образ – відображення прообразу, його властивостей і відношень на матеріальному носії. Таким носієм може бути пам'ять людини як частина нервової системи, пам'ять тварини, комп'ютера, комп'ютерних мереж;
– атомарний образ – образ деякого прообразу в цілому, без урахування його складових;
– думка – це процес і результат формування нового складеного образу шляхом комбінації існуючих.

Визначені поняття базуються на відомих поняттях в області філології, комп'ютерної лінгвістики, NLP, штучного інтелекту та узгоджуються з науковими доробками Г. Фреге, Ю. В. Крака, О. В. Бісікала, Я. О. Кохана, Ю. Р. Валькмана та ін.

Образи всього, що оточує людину і з чого вона складається (матеріально), процесів, подій і явищ будемо позначати ОбП.

На основі даних понять та використовуючи принципи об'єктно-орієнтованого моделювання із застосуванням UML була побудована модель образів світу, що характерна для сприйняття людиною.

Модель образів світу представляє собою орієнтований навантажений граф $G_{img} = V_{img} \bigcup E_{img}$, де V_{img} – множина вершин, навантаженням яких є узагальнене поняття образу, що описує основні властивості останнього; E_{img} – множина дуг, навантаженням яких є відношення «включеннями» (в) та «узагальненнями» (у) між образами (рис. 1). Так образ світу включає образ внутрішнього світу.

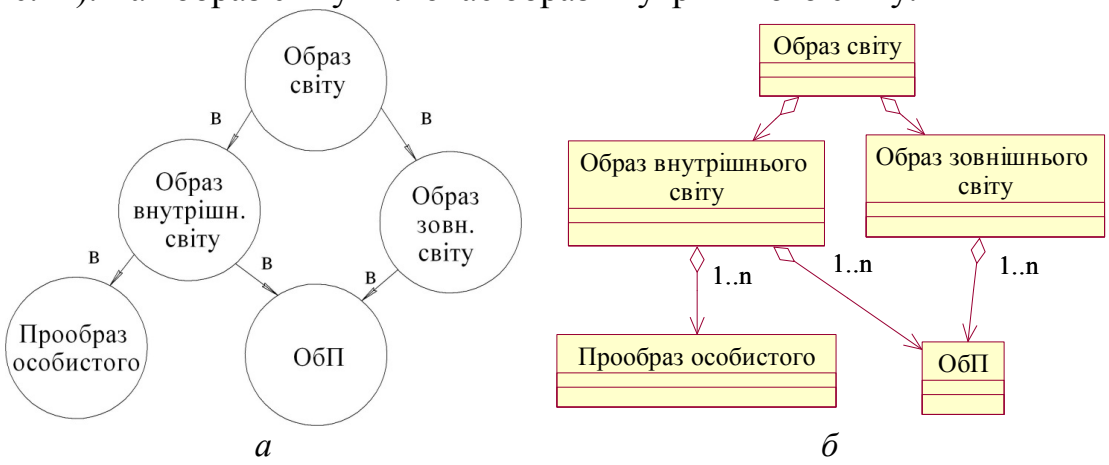


Рисунок 1 – Модель образів світу:
а – графова, б – об'єктно-орієнтована

У моделі може бути виконавець z_i – конкретна людина, спільнота тощо. Тоді для отримання моделі образів виконавця G_h^i будемо використаємо відображення $G_{img} \xrightarrow{z_i} G_h^i$. Вершинами отриманого графу є образи, які мають визначені, конкретні прообрази та відповідні властивості. Саме результат відображення відповідає поняття образу у розумінні автора.

Побудовано об'єктно-орієнтовану модель конкретизації образа прообразу, що є представленням таксономії образів людини, графове представлення якої наведено на рис. 2. Виділено властивості прообразів, які також є образами, виконана їх конкретизація.

В основу моделювання процесу формування складених образів покладено поняття думки – процесу і результату формування нового складеного образу або конструкції образів шляхом комбінації існуючих. Думка – це реалізація мети, втілена у вигляді комбінації образів з бази знань людини. Побудова думки базується на виділенні її складових – образів і відношень, формуванні складеного образу або довизначення наявного.

Людина може продукувати образні та мовні конструкції для уточнення та побудови думок. Вони мають певний смисл – образ або сукупність образів, яка відповідає думці. Слово є одним із засобів передачі думок.

Слово – комбінація звуків і / або символів і їх образів, які відповідають певному прообразу. Смислом слова є пов'язаний з ним образ особистого або загального прообразу в свідомості (пам'яті) людини. Деякі мовні одиниці не мають відповідного образу. Наприклад, заперечення «не бігати» не має образу. Його смисл формується на основі образу «бігати».

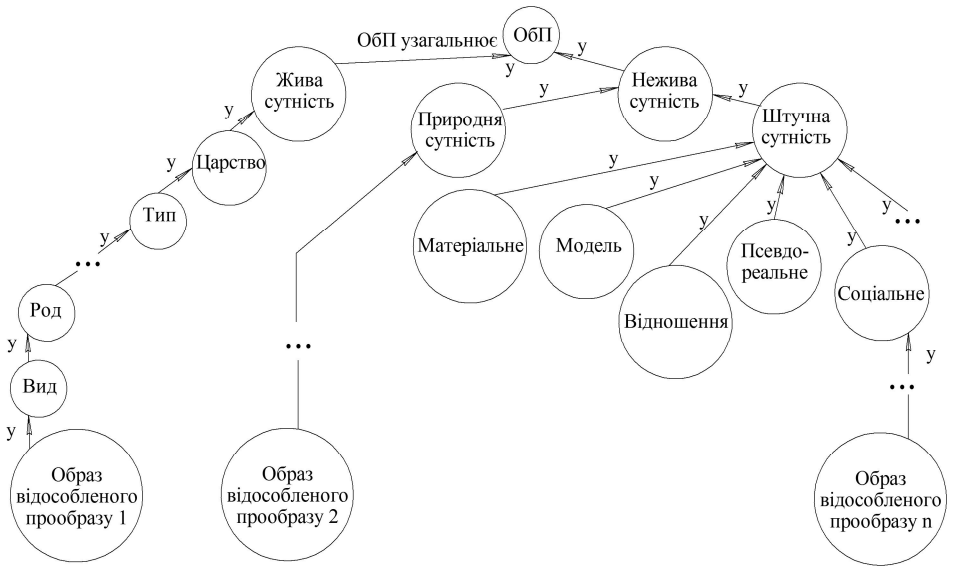


Рисунок 2 – Модель конкретизації образа прообразу

Таким чином слово є відображенням думки, смислом обох є деякий образ.

Семантична модель слова включає, а поняття семантики – об'єднує саме слово в символьному або звуковому представленні, його смисл і об'єкт дійсності, якому воно відповідає. Об'єкт дійсності визначається через компрегенсію – сукупність усіх мислимих предметів, які не є протиріччями і до яких дане слово може бути правильно застосовано. Смисл слів залежить від конкретної людини, угоди між людьми, інтонації і наголосів в усному мовленні, пунктуаційних знаків – при написанні. Крім того в смислі залишається деяка ступінь невизначеності, пов'язана з індивідуальними особливостями сприйняття світу людиною.

Думка, зазвичай, представляється більш складною мовною конструкцією (МК, словосполучення, речення). Семантична модель думки (рис. 3) є узагальненням семантичної моделі слова. Введено позначення навантаження дуг: залежність (з) та асоціація (а).

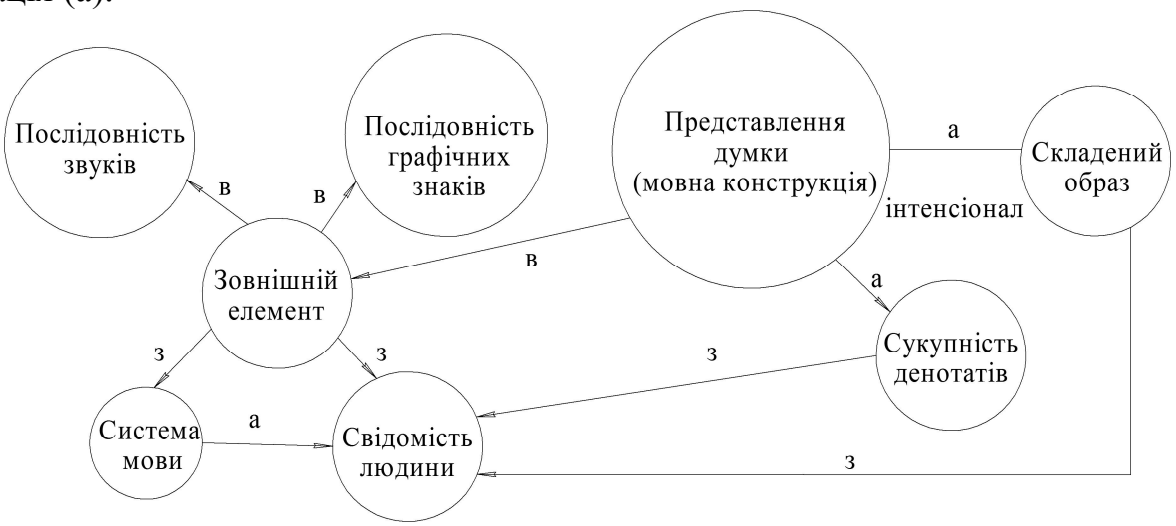


Рисунок 3 – Семантична модель думки

На основі поняття образу та створених моделей побудована конструктивно-продукційна модель мови. Для формалізації використано конструктивно-продукціне моделювання, в основі якого лежить поняття узагальненого конструктора:

$$C = \langle M, \Sigma, \Lambda \rangle, \qquad (1)$$

де M – неоднорідний носій, який включає конструктивні елементи з атрибутами та може поповнюватися, Σ – сигнатура операцій (і відповідних відношень) зв'язування, підстановки і виводу, операцій над атрибутами, Λ – множина тверджень інформаційного забезпечення конструювання (ІЗК). ІЗК (конструктивна аксіоматика) включає онтологію, мету, правила, обмеження, початкові умови та умови завершення конструювання.

Призначення конструктору полягає у формуванні множин конструкцій за допомогою операцій сигнатури, що задаються правилами ІЗК.

Для формування конструкцій необхідно виконувати ряд уточнюючих перетворень конструктору:

– спеціалізацію ($_S\mapsto$) – визначення предметної області: семантичної природу носія, кінцевої множини операцій і їх семантики, атрибутики операцій, порядку їх виконання і обмеження на правила підстановки;
– інтерпретацію ($_I\mapsto$) – зв'язування операцій сигнатури з алгоритмами виконання деякої алгоритмічної структури. При інтерпретації виконується зв'язування інформаційної моделі способу побудови конструкцій і моделі виконавця;
– конкретизацію ($_K\mapsto$) – розширення ІЗК множиною правил продукцій, завдання конкретних множин нетермінальних і термінальних символів з їх атрибутами і, при необхідності, значень атрибутів;
– реалізацію($_R\mapsto$) – послідовне виконання операції виводу для побудови конструкцій.

Конструктор образів людини має вигляд:

$$C=\langle M,\Sigma,\Lambda\rangle_S\mapsto_S C_h=\langle M_h,\Sigma_h,\Lambda_h\rangle, \qquad (2)$$

де $M_h\supset\{T\bigcup N\}$ – неоднорідний розширюваний носій, T – множина терміналів - образів, N – множина нетерміналів, Σ_h – сигнатура відношень і операцій, які виконуються на елементах носія, Λ_h – множина тверджень ІЗК.

Образ $_{\overline{w}}m_i\in M_h$ має сукупність атрибутів $\overline{w}=\{w_1,w_2,\ldots,w_n\}$. Під сукупністю будемо розуміти неоднорідну мультимножину елементів з атрибутами. Належність атрибута w_j образу m позначимо $w_j\leftarrow\!\!\lrcorner m$. Всі атрибути є образами.

Образи можуть змінюватися в часі. Кожен образ має атрибут часу створення або останньої зміни ($t\leftarrow\!\!\lrcorner m_i$). Даний атрибут змінюється в ході виконання операцій над образом і залежить від часу її виконання.

Образом світу $_tP\in M_h$ будемо називати безперервне відображення навколишнього середовища людини, представлене у вигляді динамічного потоку зображень, звуків, дотикових, смакових, нюхових і просторово-часових відчуттів, почуттів і емоцій, що відображаються нервовою системою людини під впливом матеріальних подразників. Даний образ є керованим і залежить від людини (для даної роботи це несуттєво).

Визначено такі операції над образами: конкатенація "$\cdot$", включення елементу "$\vec{\in}$", експлікація образу "$\circ,\bar{\circ}$", успадкування з уточненням "$\wedge$", успадкування з модифікацією "$\vee$", множина операцій зв'язування образів "$\Diamond_i$", узагальнення "$\uparrow$", об'єднання "$\Uparrow$", передача ">>", прийом образу"<<" , підстановка "$\rightarrow$", виведення "$\Rightarrow$", перевірка існування атрибуту "$\exists$", присвоєння ":=".

Для визначення алгоритмів виконання можливих операцій та відношень над образами виконаємо інтерпретацію конструктора (2):

$$\left\langle C_h = \left\langle M_h, \Sigma_h, \Lambda_h \right\rangle, C_A = \left\langle M_A, \Sigma_A, \Lambda_A \right\rangle \right\rangle_I \mapsto {}_{I,C_A}C_h = \left\langle M_h, \Sigma_h, \Lambda_1, Z \right\rangle, \quad (3)$$

де $M_A \supset V_A, V_A = \{A_i^0 \mid_{X_i}^{Y_i}\}$ – множина базових алгоритмів, X_i, Y_i – множини визначення та значень алгоритму $A_i^0 \mid_{X_i}^{Y_i}$, $\Lambda_1 = \Lambda_h \bigcup \Lambda_A \bigcup \Lambda_2$, $Z = \{{}_{\bar{k}}z_i\}$ – множина можливих виконавців моделі конструктора, які можуть реалізувати алгоритми C_A; $\Lambda_A = \{M_A \supset \bigcup\limits_{A_i^0 \in V_A} (X(A_i^0) \cup Y(A_i^0)) \bigcup \Omega(C_h)\}$ – неоднорідний носій, $\Omega(C_h)$ – множина конструкцій образів, які задовольняють C_h.

Виконавець ${}_{\bar{k}_i}z_i$ конструктора (3) має набір атрибутів, визначено деякі з них $\bar{k} = \{location, occupation, l\_condition, p\_characters\}$, де *location* – місце розташування (місце проживання), *occupation* – рід занять (професія, діяльність), *l_condition* – умови проживання, *p_characters* – психофізіологічні характеристики, в тому числі пов'язані зі сприйняттям і переробкою інформації.

Конструктор ${}_{I,C_A}C_h$ включає алгоритми виконання операцій: $\Lambda_2 = \{(A_1^0 \mid_{A_i, A_j}^{A_i \cdot A_j} \hookleftarrow \cdot)$, $(A_2^0 \mid_S^{A_1} \hookleftarrow :)$, $(A_3 \mid_{m_1, m_2, P}^{m_1 \cdot m_2} \hookleftarrow \cdot)$, $A_4 \mid_{m, \bar{m}}^{\bar{m}} \hookleftarrow \vec{\in})$, $(A_5 \mid_{m_1, m_2}^{m^*} \hookleftarrow \circ)$, $(A_6 \mid_{m_1, m_2}^{m^*} \hookleftarrow \bar{\circ})$, $(A_7 \mid_{m_1, m_2}^{m^*} \hookleftarrow \wedge)$, $(A_8 \mid_{m_1, m_2, m_3}^{m^*} \hookleftarrow \vee)$, $(A_9 \mid_{m_i, m_j}^{m'} \hookleftarrow \Diamond)$, $(A_{10} \mid_{\bar{m}}^{\bar{m}} \hookleftarrow \uparrow)$, $(A_{11} \mid_{\bar{m}, m}^{\bar{m}} \hookleftarrow \Uparrow)$, $(A_{12} \mid_{m, P}^{P} \hookleftarrow_h >>)$, $(A_{13} \mid_{m, P}^{P} \hookleftarrow_s >>)$, $(A_{14} \mid_P^m \hookleftarrow_h <<)$, $(A_{15} \mid_P^m \hookleftarrow_s <<)$, $(A_{16} \mid_{m_1, m_2}^{c} \hookleftarrow \exists)$, $(A_{17} \mid_{l_h, l_q, f_i}^{f_i} \hookleftarrow \Rightarrow)$, $(A_{18} \mid_{f_i, \Psi}^{f_j} \hookleftarrow \mid\Rightarrow)$, $(A_{19} \mid_{\sigma, \Psi}^{\bar{\Omega}} \hookleftarrow \|\Rightarrow)$, $(A_{20} \mid_{a, b}^{b} \hookleftarrow :=)\}$.

Для уточнення введених операцій виконано конкретизацію конструктора (3):

$$ {}_{I,C_A}C_h = \left\langle M_h, \Sigma_h, \Lambda_1, Z \right\rangle \; {}_K \mapsto {}_{K,I,C_A}C_h = \left\langle M_h, \Sigma_h, \Lambda_2, Z \right\rangle, \quad (4)$$

де $\Lambda_2 = \Lambda_1 \bigcup \Lambda_3$, $\Lambda_3 \supset \{M_h \supset T \bigcup N$, $T = \{K, P, K_{pw}, K_s, K_{aw}\}\}$ – множина терміналів, K – конструкція у вигляді сукупності образів, K_s – конструкція впорядкованих образів звуків, K_{pw} – конструкція впорядкованих образів символів писемної мовної конструкції (МК), K_{aw} – конструкція образів, отриманих при спостереженні дій, погляду, міміки і т.п., $N = \{\sigma, \eta, \alpha, \beta, \delta, \chi, \gamma, \kappa\, \mu, \theta, \nu, \lambda\}$ – множина нетерміналів, σ – початковий нетермінал.

Наведемо кілька правил. Правила підстановки $s_1 - s_3$ дозволяють сформувати новий образ на основі операції експлікації:

$$s_1 = \left\langle \sigma \to K \right\rangle, \; s_2 = \left\langle K \to \vec{\in}(K, \chi) \right\rangle, \; s_3 = \left\langle \chi \to \circ(P, \varepsilon) \mid \circ(K, P) \mid \circ(P, K) \right\rangle. \quad (5)$$

Правила підстановки $s_4 - s_5$ дозволяють виконати успадкування образу з уточненням:

$$s_4 = \left\langle K \to \vec{\in}(K, \wedge(\chi, \gamma)) \right\rangle, \; s_5 = \left\langle \gamma \to \circ(P, K) \mid \circ(K, P), K \to \vec{\in}(K, \gamma) \right\rangle. \quad (6)$$

Результатом реалізації конструктора (4) є множина образів (МК) в цілому і їх частин ${}_t\Omega(C_h({}_{\bar{k}_i}z_i))$: ${}_t\Omega(C_h({}_{\bar{k}_i}z_i)) \supset ({}_t\Omega_s(C_h({}_{\bar{k}_i}z_i)) \bigcup {}_t\Omega_{pw}(C_h({}_{\bar{k}_i}z_i)) \bigcup \bigcup {}_t\Omega_{aw}(C_h({}_{\bar{k}_i}z_i)))$, де множини ${}_t\Omega(C_h({}_{\bar{k}_i}z_i))$ – всіх сформованих образів виконавцем ${}_{\bar{k}_i}z_i$ на момент часу t,

${}_t\Omega_{pw}(C_h(z_i))$ – всіх образів писемних МК, ${}_t\Omega_s(C_h(z_i))$ – мовленнєвих конструкцій (в тому числі тих, які відповідають ${}_t\Omega_{pw}(C_h(z_i)))$, ${}_t\Omega_{aw}(C_h(z_i))$ – інші образи МК.

Конструкції ${}_t\Omega(C_h({}_{\bar{k}_i}z_i))$ – елементи спілкування, притаманні конкретному виконавцю ${}_{\bar{k}}z_i \in Z$, будемо називати індивідуальною мовою. Вільна мова – це множина потенційно можливих конструкцій, які виконавець (людина) може розпізнавати (розуміти) і використовувати для передачі інформації.

Нехай існує деякий підмножина $\overline{Z} \subseteq Z$ з n виконавців конструктора C_h. Мову спільноти виконавців $\overline{Z}$ розглядатимемо як сукупність конструкцій, побудовану на носії конструктора (4) в результаті його реалізації $L(t) = \bigcup_b ({}_t\Omega^*(C_h({}_{\bar{k}_i}z_i)) \bigcap {}_t\Omega^*(C_h({}_{\bar{k}_j}z_j)))$, де $b = ({}_{\bar{k}_i}z_i, {}_{\bar{k}_j}z_j \in \overline{Z}, {}_{\bar{k}_i}z_i \neq {}_{\bar{k}_j}z_j)$, ${}_t\Omega^*(C_h({}_{\bar{k}_i}z_i)) = {}_t\Omega_{pw}(C_h({}_{\bar{k}_i}z_i)) \bigcup {}_t\Omega_s(C_h({}_{\bar{k}_i}z_i)) \bigcup \Omega_{aw}(C_h({}_{\bar{k}_i}z_i))$. Мова існує в деякий момент часу t. МК належить мові, якщо існує два і більше її носія, здатних її приймати і передавати $({}_t\Omega^*(C_h({}_{\bar{k}}z_i)) \bigcap {}_t\Omega^*(C_h({}_{\bar{k}}z_j)) \neq \varnothing, {}_{\bar{k}_i}z_i \neq {}_{\bar{k}_j}z_j)$. Письмова мова спільноти виконавців $L_{pw}(t) = \bigcup_b ({}_t\Omega_{pw}(C_h({}_{\bar{k}_i}z_i)) \bigcap {}_t\Omega_{pw}(C_h({}_{\bar{k}_j}z_j)))$; усна – $L_s(t) = \bigcup_b ({}_t\Omega_s(C_h({}_{\bar{k}_i}z_i)) \bigcap {}_t\Omega_s(C_h({}_{\bar{k}_j}z_j)))$.

Використовуючи апарат конструкторів побудована модель мовної конструкції тексту, що відображає його лексичну та синтаксичну складову.

Для вирішення задачі зіставлення текстів побудована модель графового представлення тексту:

$$C = \langle M, \Sigma, \Lambda \rangle \; {}_S\mapsto C_g = \langle M_g, \Sigma_g, \Lambda_g \rangle, \tag{7}$$

де M_g – розширюваний носій, що включає множини конструкцій-графів, МК і їх елементів, Σ_g – множина операцій і відношень на елементах M_g, Λ_g – ІЗК.

Носій включає множини термінальних і нетермінальних елементів $M_g \supset T_g \bigcup N_g$. Терміналами є мовні конструкції, побудовані конструктором ${}_A C_T$ і їх складові (T_T), а також конструкції графів і їх складових $T_g = \overline{\Omega} \bigcup \Omega_g \bigcup T_T \bigcup V \bigcup E$, Ω_g – множина конструкцій-графів, V, E – множин вершин і дуг з їх атрибутами.

Вершина має атрибути $\overline{w}_v = \langle id, content, tokens \rangle$, id – ідентифікатор, приймає цілочислені значення, $content$ – частина текстової конструкції, $tokens$ – список, що містить ознаки початку мовних конструкцій. Атрибути дуги $\overline{w}_e = \langle id, routes, start, end \rangle$, id – ідентифікатор, приймає цілочислені значення, $routes$ – множина номерів шляхів, в які входить дуга (вказує на порядок обходу графа), *start, end* – вершини, які є інцедентними до дуги *e*.

Навантажений граф будемо позначати як ${}_{\overline{w}_g}G = \langle V, E \rangle$, $V = \{{}_{\overline{w}_{vi}}v_i\}, E = \{{}_{\overline{w}_{ej}}e_j\}$ – множини вершин і дуг, навантажених атрибутами. Кожна множини містить порожній елемент.

Граф має такі атрибути $\overline{w}_g = \langle start\_v, last\_v, current\_v, amount\_l \rangle$, де *start_v* – стартова вершина графа, *last_v* – остання додана вершина, *current_v* – поточна вер-

шина при формуванні графа, *amount_l* – кількість циклів, в які входить стартова вершина.

Для побудови графа виконано інтерпретацію та конкретизація конструктора (7), та визначено множину правил, загальний вид яких

$$\psi_i = \langle s_i, g_i \rangle,\ s_i = \langle \bar{s}_i, \tilde{s}_i \rangle,\ g_i = \langle \bar{g}_i, \tilde{g}_i \rangle, \tag{8}$$

де $\bar{s}_i$, $\tilde{s}_i$ – відношення підстановки для розпізнавання мовної конструкції і побудови конструкції графа відповідно, $\bar{g}_i$, $\tilde{g}_i$ – операції над атрибутами мовної конструкції і графа, його вершин і дуг відповідно. У разі якщо операції над атрибутами не виконуються, відношення підстановки має вигляд $\psi = \langle s, \varepsilon \rangle$.

Для зменшення ресурсів, необхідних для порівняння текстів, створено конструктор для стиснення графів:

$$C = \langle M, \Sigma, \Lambda \rangle_S \mapsto C_C = \langle M_C, \Sigma_C, \Lambda_C \rangle, \tag{9}$$

де $\Lambda_C = \{M_C = V \bigcup E \bigcup \Omega_g,\ \Sigma_C = \{\{\cdot, :=\}, \{\rightarrow, |\Rightarrow, \|\tilde{\Rightarrow}\} \bigcup \Psi_C\},\ \Psi_C = \{\psi_C = \langle s_i, g_i \rangle\}\}$.

Виконано конкретизацію конструктора для стиснення графа:

$$C_{C\,K} \mapsto C_{CK} = \langle M_{CK}, \Sigma_{CK}, \Lambda_{CK} \rangle, \tag{10}$$

де $M_{CK} = M_C \bigcup N_{CK}, N_{CK} = \{\sigma\}$.

Правила, що дозволяють стиснути граф:

$$\begin{gathered} s_1 = \langle \sigma \rightarrow G\sigma \rangle,\ s_2 = \langle G\sigma_d \rightarrow G\sigma \rangle, \\ g_1 = g_2 = \langle e_1 := (v_i, v_j, G), e_2 := (v_j, v_k, G), \\ \div (routes \lrcorner e_1 = routes \lrcorner e_2, 4, d := true, content \lrcorner v_i := \cdot (content \lrcorner v_i, content \lrcorner v_j), \\ tokens \lrcorner v_i := \cdot (tokens \lrcorner v_i, tokens \lrcorner v_j), end \lrcorner e_1 := v_k) \rangle,\ s_3 = \langle \sigma \rightarrow \varepsilon \rangle. \end{gathered} \tag{11}$$

Реалізація інтерпретованого конструктора ${}_A C_C$ полягає в формуванні графових конструкцій, які мають однозначну відповідність конструкціям $\overline{\Omega}({}_A C_T)$ і $\overline{\Omega}({}_A C_g)$ шляхом виконання алгоритмів, пов'язаних з операціями сигнатури, за правилами аксіоматики:

$${}_A C_{C\ R} \mapsto \overline{\Omega}({}_A C_C), \overline{\Omega}({}_A C_C) \subset \Omega({}_A C_C). \tag{12}$$

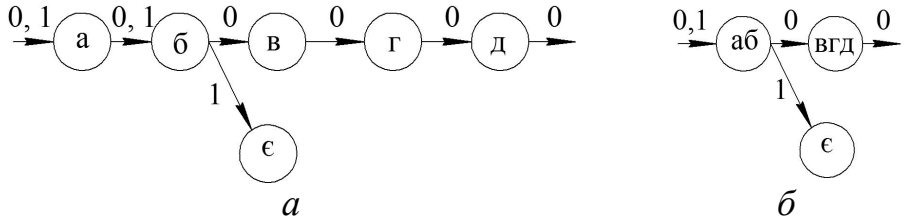
Для зменшення розміру графу та прискорення його обробки пропонується метод стиснення графа з текстовим навантеженням. Даний метод базується на:

– операціях гомеоморфізму: стиснення графа за парою суміжних ребер шляхом виключення вершини, стягування ребер;
– алгоритму групування вузлів графа за атрибутами (SNAP);
– принципах обробки соціальних графів (збереження базової частини та відмінностей).

Метод полягає у стисненні графа за парами суміжних дуг шляхом виключення вершини та збільшення навантаження попередньої вершини, за умови рівності навантаження дуг (рис. 4). Стиснення виконується для всіх пар дуг, що мають однакові навантаження та не є інцидентними до стартової.

Графового представлення тексту та методу стиснення графа дозволило розробити метод порівняння структурованих документів. Під структурованими документами будемо розуміти електронні документи, представлені файлами у форматі doc/docx, які мають логічну структуру за змістом (розділи та підрозділи, порядок

яких має значення) та відповідне форматування. Під процесом порівняння структурованого документа будемо розуміти впорядковану послідовність дій, що виконуються над документом, поданим для перевірки на наявність запозичень, та базою структурованих документів, з якою виконується порівняння. Дії виконуються за етапами, визначеними в роботі.



а *б*

Рисунок 4 – Приклад стиснення фрагмента графа:
а – фрагмент до стиснення; б – фрагмент після стиснення

Для формалізації методу побудовано конструктор процесу порівняння:

$$C = \langle M, \Sigma, \Lambda \rangle \ {}_s\!\mapsto C_{CP} = \langle M_{CP}, \Sigma_{CP}, \Lambda_{CP} \rangle, \tag{13}$$

де M_{CP} – носій, елементами якого є алгоритми виконання операцій над документом $\{D_i^0\}$ та його атрибутами й алгоритми над текстом $\{D_i^1\}$ (визначені у роботі), Σ_{CP} – сигнатура операцій та відношень для побудови конструкцій – алгоритмів і сценаріїв, Λ_{CP} – множина тверджень ІЗК.

В роботі виконано ряд уточнюючих перетворень C_{CP}. Визначено правила підстановок, які дозволяють формалізувати процес порівняння структурованих документів відповідно до визначених етапів. Наведемо частину правил для роботи зі структурою документа

$$\begin{aligned} s_1 &= \left\langle CP \,|^{marks, ul\_marks}_{doc, template} \rightarrow D_1^0 \,|^{sections}_{doc} \cdot \beta \,|^{number}_{section, template} \right\rangle, \\ s_2 &= \left\langle \beta \,|^{number}_{section, template} \rightarrow D_2^0 \,|^{number}_{section, template} \cdot \beta \,|^{number}_{section, template} \right\rangle, \\ s_4 &= \left\langle \beta \,|^{number}_{section, template} \rightarrow D_2^0 \,|^{number}_{section, template} \cdot \gamma \,|^{number}_{section, template} \right\rangle, \end{aligned} \tag{14}$$

де $count\_number = 0$ – ознака того, що для розділу не знайдено відповідника у шаблоні. Правило s_1 реалізує побудову структурного дерева *sections* для документа *doc*. Правила s_2, s_4 – ідентифікація структурних елементів документа відповідно до шаблона template.

Пошук у базі документів *db* структурних елементів, з якими може бути проведене порівняння (D_4), і попередня обробка тексту структурних елементів D_1^1

$$s_6 = \left\langle \gamma \,|^{text', db\_texts}_{sections, template} \rightarrow D_4^0 \,|^{s\_texts, db\_texts}_{sections, db} \cdot D_1^1 \,|^{text'}_{s\_texts_i} \cdot \delta \,|^{ul\_mark}_{text', db\_texts} \right\rangle. \tag{15}$$

Алгоритми $\{D_i^0\}, \{D_j^1\}$ та модель (13) – (15) складають метод порівняння структурованих документів, схема якого наведена у роботі.

Третій розділ присвячено експериментальній частині дослідження. Виконано аналіз та виявлення подібності мовних конструкцій та структурованих документів на основі конструктивної моделі природної мови (2) – (6). Виконано моделювання текстових фрагментів однієї тематики, проте з різною лексично-синтаксичною структурою та різними природними мовами. Для наочності виконано графове представлення текстів, проведено аналіз структури та взаємозв'язків виділених образів. Для структурованих документів визначено їх образне представлення, за яким виконано

їх зіставлення. В рамках експерименту структура визначалася вимогами до документів, які є пояснювальними записками до дипломних проектів.

За результатами можна зробити висновок, що зміна лексики та корегування структури має незначний вплив на образне представлення текстів та документів, а тому даний підхід може бути використано для виявлення семантичних запозичень.

Для перевірки можливості застосування моделі графового представлення текстів для виявлення запозичень в умовах маскування розроблена модель процесів маскування, що передбачають побудову та застосування сценаріїв запозичень. Модель складається з двох конструкторів: сценаріїв та модифікації текстів. Перший конструктор має вигляд

$$C = \langle M, \Sigma, \Lambda \rangle_s \mapsto C_{sc} = \langle M_{SC}, \Sigma_{SC}, \Lambda_{SC} \rangle, \quad (16)$$

де M_{SC} – розширюваний носій, елементами якого є термінали та нетермінали, Σ_{SC} – сигнатура операцій та відношень для побудови конструкцій – алгоритмів і сценаріїв, Λ_{SC} – множина тверджень ІЗК, які будуть розглянуті далі. $M_{SC} \supset T \bigcup N$. Термінали ($T$) включають кінцеву множину алгоритмів виконання операцій безпосередньо над текстом $\{B_i^0\}$ і додаткових (допоміжних) $\{B_i^1\}$, які визначено у роботі, а також сконструйовані алгоритми і їх послідовності – сценарії та необхідні для них дані. Базові алгоритми реалізують атомарні (умовно неподільні) дії. Нетермінали – допоміжні елементи, складові множини абстрактних алгоритмів.

Текст, до якого застосовуються базові алгоритми, має ряд визначених атрибутів.

Конструктор модифікації текстів:

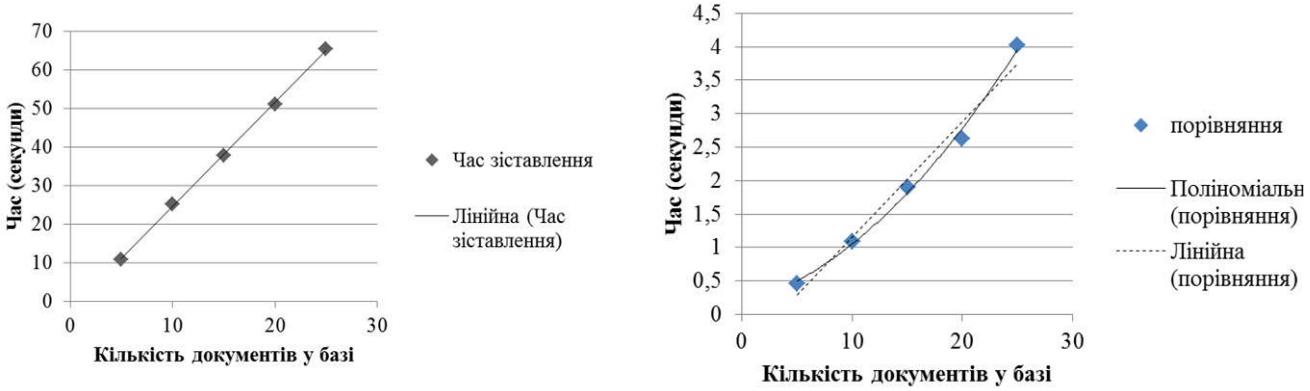
$$C = \langle \mathrm{M}, \Sigma, \Lambda \rangle_S \mapsto C_{AT} = \langle \mathrm{M}_{AT}, \Sigma_{AT}, \Lambda_{AT} \rangle, \quad (17)$$

де $\mathrm{M}_{AT}, \Sigma_{AT}, \Lambda_{AT}$ – носій, сигнатура та множина тверджень ІЗК відповідно. Елементами носія є термінали і нетермінали. Термінали включають кінцеву множину сконструйованих сценаріїв $\Omega(C_{A,SC})$, а також дані для їх застосування, тексти і їх елементи: символи електронного подання текстів та мовні конструкції: слова, речення, абзаци. Нетермінали – допоміжні елементи.

Таким чином, розроблений формалізм утворив теоретичну базу для створення автоматизованої системи складання тестів для систем анти плагіату.

Проведено комп'ютерні експерименти зі стиснення графів, які показали, що розроблений метод дозволяє заощадити близько 97% пам'яті, необхідної для зберігання графового представлення текстів та спрощує структуру графу на 90 – 97% за різними показниками. Області переваги стиснутого представлення складає 100%.

Проведено комп'ютерні експерименти для дослідження часової ефективності операцій побудови графового представлення та зіставлення текстів та структурованих документів. Для останніх виділено чотири етапи: попередня обробка, побудова графового представлення, порівняння та оцінка результату. За регресійним аналізом встановлено лінійну залежність часової ефективності операції зіставлення структурованих документів (текстів) від їх розміру та розміру бази та близьку до лінійної – за складовим операції (рис. 5) Аналіз складових операції зіставлення показав, що основний час (близько 94%) витрачається на отримання наборів графів. Середній час зіставлення структурованого документу (середній розмір 172 тис. знаків) з всіма наявними у базі роботами складає від 11 до 65 секунд при базі від 0,6 до 3,8 млн. знаків.



а) *б)*

Рисунок 5 – Залежність часу зіставлення від розміру бази: а – загальний, у тому числі б – порівняння

Підготовка та проведення експерименту з оцінки функціональної ефективності розробленого ПЗ дозволила визначити ряд факторів, що значно зменшують кількість аналогів, доступних для порівняння. Проведені експерименти показали, що розбіжність у результатах роботи розробленої програми та аналогу не перевищує 5%.

Аналіз текстів та виявлених запозичених фрагментів показав, що розбіжності результатів зіставлення зумовлені:

– різною інтерпретацією поняття слова;
– різницею у кількості слів в документах після попередньої обробки аналогом та розробленою програмою;
– впливом форматування та невизначеністю впливу тексту колонтитулів.

Виконано комп'ютерний експеримент для визначення ступеню чутливості комп'ютерної реалізації моделі графового представлення тексту до впливу сценаріїв маскування запозичень.

Під ступенем чутливості будемо розуміти

$$S_d = \frac{1}{N}\sum_{i=1}^{N}\frac{Per(Original)_i - Per(Disguise)_i}{\max(Per(Original)_i; Per(Disguise)_i)}, \quad (18)$$

де *Per(Original), Per(Disguise)* – відсоток запозичень у тексті, поданому на перевірку, без та з маскуваннями відповідно, *N* – кількість пар текстів, поданих на перевірку. Дана величина вказує на те, у скільки разів маскування знизило відсоток запозичень.

Визначено у скількох відсотках випадках відбулося підвищення оригінальності за рахунок маскувань, розрахувавши даний показник за формулою:

$$R_d = \frac{1}{N}\sum_{i=1}^{N} sign(Per(Original)_i - Per(Disguise)_i)\cdot 100\%. \quad (19)$$

Визначено середній показник зміни відсотку запозичень як

$$A_d = \sum_{i}^{N} Per(Original)_i - Per(Disguise)_i / N. \quad (20)$$

Середній показник S_d склав $4{,}7\cdot 10^{-5}$ – ступінь підвищення оригінальності текстів за рахунок маскувань, що спостерігається в близько 40% випадків. Середнє зниження відсотку плагіату становить близько 0,007%, що є не суттєвим.

У четвертому розділі представлено результати розробки об'єктно-орієнтованих моделей графового представлення текстів, документів та її комп'ютерної (програмної) реалізації для виявлення запозичень у текстах та струк-

турованих текстових документах. Встановлено зв'язок алгоритмічної складової моделі графового представлення тексту та логіки програми.

Представлено результати розробки трьох додатків. Перший – для автоматизованого формування тестів на основі конструктивної моделі сценаріїв модифікації.

Додаток генерує тестові набори даних для оцінки здатності програм-антиплагіатів демаскувати запозичення. Для користування функціоналом розробленої програми реалізовано інтерфейс користувача, зі структурою діалогу на базі екранних форм. Форма інтерфейсу користувача має дві вкладки – для ручного та випадкового конструювання сценаріїв (рис. 6), що дозволяє обирати кількість елементів сценаріїв, їх суть та значення вхідних параметрів, створюючи розмаїття сценаріїв. Встановлено зв'язок алгоритмічної складової моделі процесів маскувань та логіки програми.

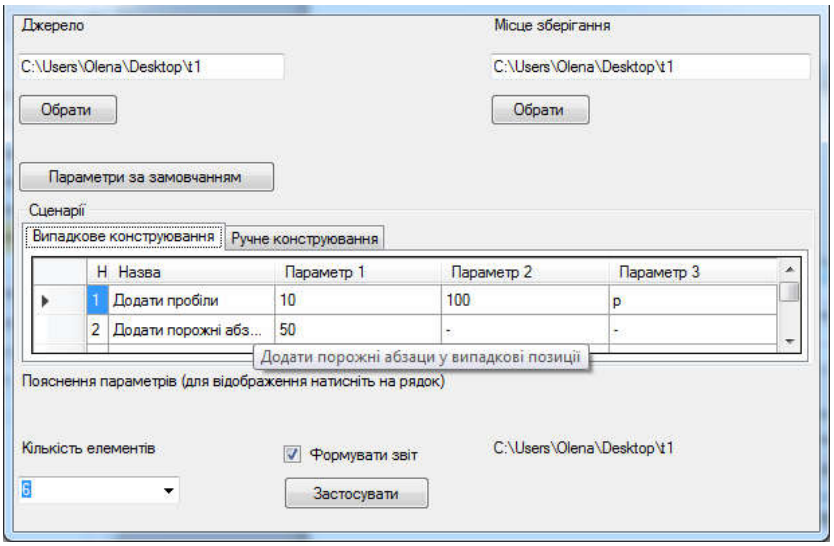


Рисунок 6 – Форма конструювання сценарію

Другий додаток дозволяє виконувати порівняння структурованих документів з базою (рис. 7) та визначати відсоток запозичень за розділами та документом в цілому. Порівняння ведеться згідно розробленого структурного шаблона, представленого xml-файлом. Для розділів, що не були розпізнані за шаблоном, можна задати відповідність у ручному режимі.

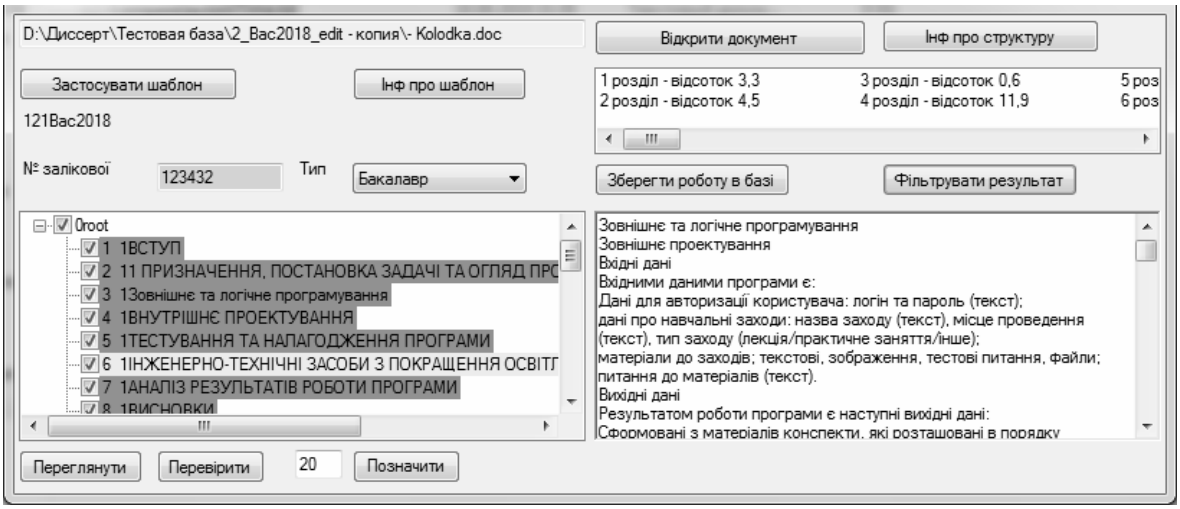


Рисунок 7 – Головне вікно програми для порівняння структурованих документів

Третій додаток дозволяє: порівнювати два документи з визначенням загального відсотку запозичень та за фрагментами зі збереженням результатів; фільтрувати

результати порівняння: задати мінімальну допустимі довжину запозиченого фрагменту та відстань між фрагментами, що спрощує структурний аналіз результату перевірки; порівняти декілька текстових документів з декількома (рис. 8), визначивши загальний відсоток запозичень та їх структурний склад за кожною парою документів.

Встановлено зв'язок алгоритмічної складової моделі графового представлення тексту та логіки програми.

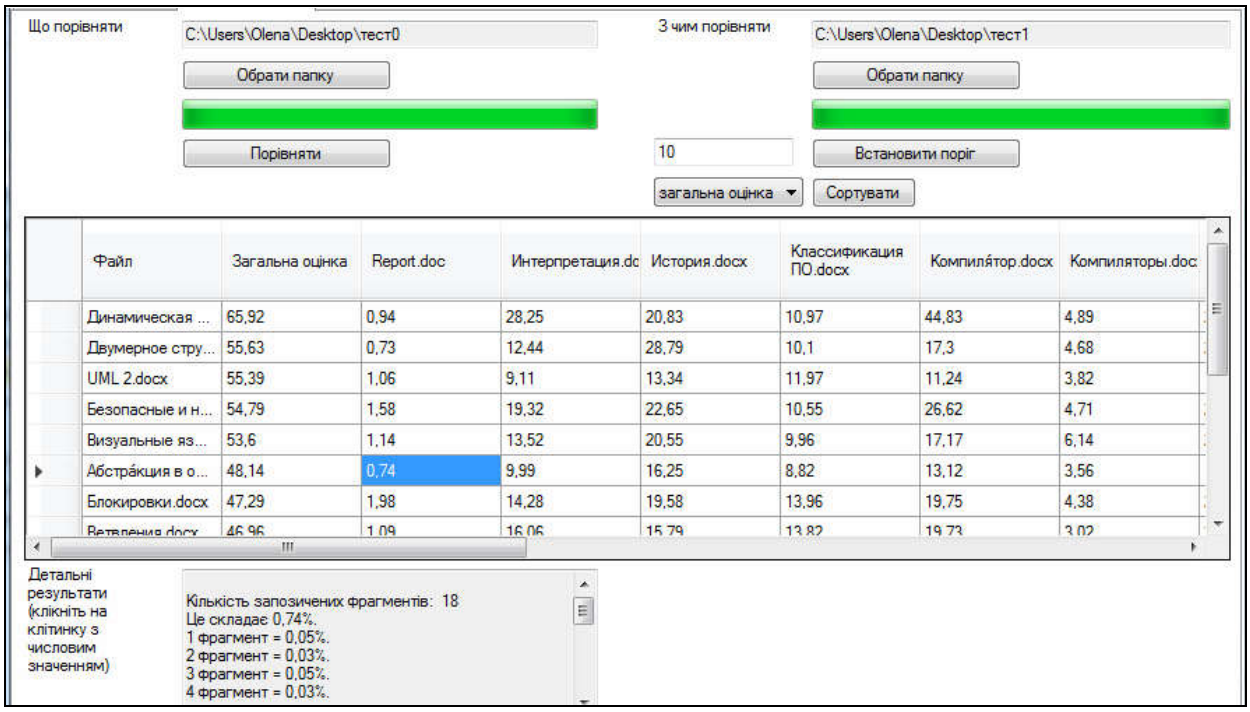


Рисунок 8 – Результати порівняння декількох файлів

ВИСНОВКИ

У дисертаційній роботі розроблено конструктивно-продукційні та об’єктно-орієнтовані моделі природної мови та тексту, структурованого документу, методи та засоби зіставлення документів та маскувань запозичень, які дозволили вирішити науково-технічну задачу виявлення запозичень у структурованих документах кваліфікаційного характеру.

Дані моделі та засоби утворюють єдиний комплекс (рис. 9), що охоплює:

– мову та мовлення – формування мовних конструкцій (компоненти 1, 2, 3, 5, 6);
– семантичні, лексичні, синтаксичні складові мовних конструкцій (2, 5, 6);
– процеси маскування (7,10) та виявлення запозичень (4, 8, 11) з урахуванням структури документів (9, 12, 13).

За результатами дисертаційного дослідження опубліковано 24 наукові праці: 7 статей (в тому числі 2 у виданнях, що індексуються Scopus і Web of Science), 2 авторські свідоцтва та 15 робіт апробаційного характеру.

Основні наукові та практичні результати:

1) за проведеним у роботі аналізом мовних моделей встановлено, що більшість з них не враховує виконавця та його особливості, ігнорують процеси мислення людини, враховують не всі аспекти мови. Методи, алгоритми та засоби обробки мовних конструкцій (текстів) для виявлення запозичень не враховують структуру документів; деякі є чутливими до механізмів маскування запозичень. Виявлено потребу розробки єдиного підходу до моделювання лексики,

синтаксису, семантики мови, її конструкцій та засобів їх порівняння з метою виявлення запозичень;

2) вперше виконано формалізацію процесів формування образів людини засобами об’єктно-орієнтованого моделювання, побудовано ієрархію образів на основі спільності атрибутів, що дозволило представити смисл слова і відобразити його зв’язок з реальними речами в рамках поняття «семантика слова» і використано при побудові конструктивно-продукційної моделі мови;
3) вперше розроблено конструктивно-продукційну модель природної мови на основі образного представлення дійсності, використання якої дає змогу зменшити вплив лексичних та синтаксичний змін на семантику текстів в задачах виявлення запозичень;
4) вперше розроблено моделі (конструктивно-продукційну та об’єктно-орієнтовану) мовних конструкцій та їх графового представлення, метод і алгоритми зіставлення текстів структурованих документів. Отримали подальший розвиток методи і засоби конструктивно-продукційного моделювання: встановлено зв'язок конструктивно-продукційних моделей з об’єктно-орієнтованими. Їх комп’ютерна реалізація дозволяє виявляти запозичення у структурованих документах при зміні порядку лексем з прийнятними показниками часової ефективності операції порівняння;
5) в рамках подальшого розвитку методів обробки графів розроблено метод стиснення графа з текстовим навантаженням для підвищення ефективності комп’ютерної реалізації моделі графового представлення тексту, що дозволило застосувати механізм серіалізації об’єктів при формуванні бази структурованих документів;
6) вперше розроблено конструктивно-продукційну та об’єктно-орієнтовану моделі процесів маскування запозичень, що дозволило автоматизувати процес формування тестів для перевірки здатності демаскування запозичень у програмах-антиплагіатах;
7) на єдиній теоретичній основі розроблено комплекс моделей мови, мовних конструкцій та засоби їх обробки, який дозволяє вирішувати задачу виявлення запозичень та узгоджуються з відомими теоріями та методами моделювання та обробки текстів і є подальшим розвитком методів виявлення семантичних запозичень;
8) виконано комп’ютерні експерименти для дослідження часової та функціональної ефективності використання розробленої моделі текстів в задачах виявлення запозичень. Отримані результати за кількістю виявлених запозичень були порівняні з аналогом – різниця не перевищує 5%. Середній час зіставлення структурованого документу складає від 11 до 65 секунд при базі від 0,6 до 3,8 млн. знаків і має лінійну залежність від розміру текстів. Показники є прийнятними і роблять доцільним впровадження та подальше використання програмної реалізації моделі в академічному середовищі;
9) розроблене програмне забезпечення дозволяє сприяти виконанню вимог ЗУ «Про вищу освіту» щодо виявлення академічного плагіату у роботах здобувачів вищої освіти та впроваджене в Дніпровському національному університеті залізничного транспорту ім. академіка В. Лазаряна. Результати дисертаційному роботи впроваджено у програмних проектах філії ПКБ ІТ АТ «Укрзалізниця» та ТОВ «СОВЛАНУТ».

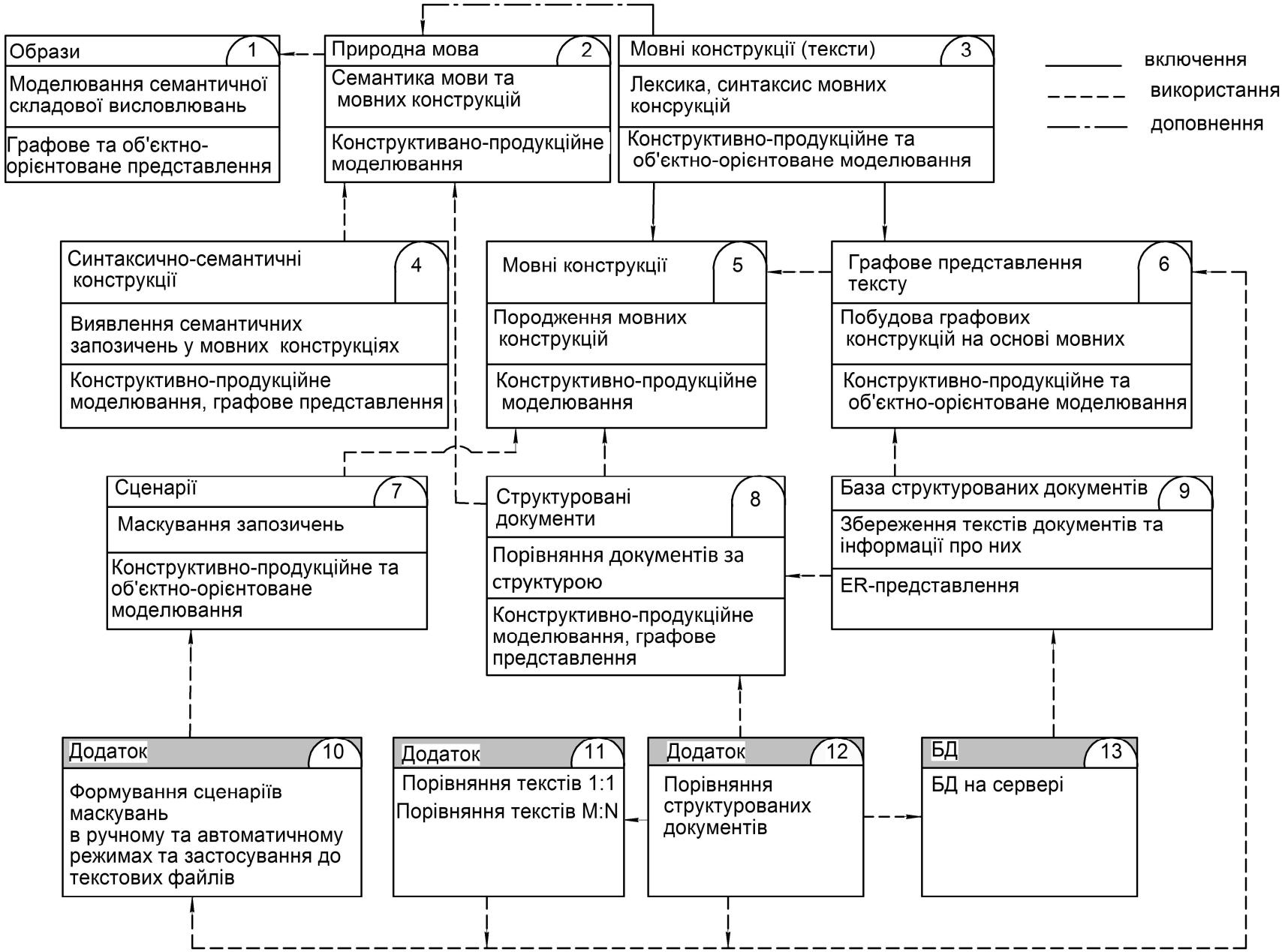


Рисунок 9 – Зв'язок моделей

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні результати дисертації:

1. Шинкаренко В. І. Система контролю плагіату в студентських роботах / В. І. Шинкаренко, О. С. Куроп'ятник // Восточно-Европейский журнал передовых технологий //научный журнал. – Харьков: Технологический центр, 2012 – № 4/2 (58). – с. 32 – 36. *Видання включено до НМБ Index Copernicus International.*

2. Шинкаренко В. И. Объектно-ориентированная модель смысловых составляющих языковых конструкций / В. И. Шинкаренко, Е. С. Куропятник. – Искусственный интеллект // науч. журнал. – НАН Украины, Институт проблем искусственного интеллекта. – № 4. – 2014. – с. 44 – 56

3. Шинкаренко В. И. Конструктивно-продукционная модель графового представления текста / В. И. Шинкаренко, Е. С. Куропятник – Проблемы программирования // науч. журнал. – № 2 – 3. – 2015. – с. 63 – 72. Передрук: Shynkarenko V. Constructive-synthesizing model of text graph representation // V. Shynkarenko, O. Kuropiatnyk – CEUR Workshop Proceedings. – 2016. – Vol. 1631. – P. 63 – 72. *Видання включено до НМБ Scopus.*

4. Шинкаренко, В. И. Проблемы выявления плагиата и анализ инструментального программного обеспечения для их решения / В. И. Шинкаренко, Е. С. Куропятник // Наука та прогрес транспорту. — 2017. — № 1 (67). — С. 131—142. — doi: 10.15802/stp2017/94034. *Видання включено до НМБ Index Copernicus International.*

5. Shynkarenko V. Constructive Model of the Natural Language / V. Shynkarenko, O. Kuropiatnyk // ActaCybernetica. – 2018 – Vol. 23, Nr 4. – P. 995-1015. – DOI: 10.14232/actacyb.23.4.2018.2. *Видання включено до НМБ Scopus, Web of Science.*

6. Шинкаренко В. Формирование тестов для проверки способности демаскировки заимствований в программах выявления плагиата / В. Шинкаренко, Е. Куропятник // Information Technologies & Knowledge. – 2018. – Vol.12, Nr 1. – P. 84 – 100.

7. Куроп'ятник О. С. Конструктивне та об'єктно-орієнтоване моделювання текстів для виявлення запозичень / О. С. Куроп'ятник // Системні технології. Регіональний міжвузівський збірник наукових праць. – Випуск 4 (123). – Дніпро, 2019. – с. 34 – 47.*Видання включено до НМБ Index Copernicus International.*

Опубліковані праці апробаційного характеру:

8. Шинкаренко В. І., Куроп'ятник О. С. Система виявлення плагіату в студентських роботах / В. І. Шинкаренко, О.С. Куроп'ятник // Сучасні інформаційні технології на транспорті, в промисловості та освіті : міжнарод. наук.-практич. конф., 5-6 квітня 2012 р.: тези доп. – Д.: ДНУЗТ, – 2012. – С. 124.

9. Куроп'ятник О. С. Конструювання текстових структур в задачах запобігання плагіату / О.С. Куроп'ятник // Сучасні інформаційні технології на транспорті, в промисловості та освіті : міжнарод. наук.-практич. конф., 18-19 квітня 2013 р.: тези доп. – Д.: Вид-во ДНУЗТ, – 2013. – с. 70.

10. Шинкаренко В. И., Куропятник Е. С. Моделирование образного представления действительности средствами конструктивно-продукционных структур / В. И. Шинкаренко, Е. С. Куропятник // Информационные технологии в металлургии и машиностроении: междунар. научнотех. конф., 24 – 26 марта 2015 г.: тезисы докл. – Д.: НМетАУ, – 2015. – с. 80.
11. Шинкаренко В. И., Куропятник Е. С. Моделирование образного представления экономического обеспечения транспорта // Проблемы экономики транспорта: XIII междунар. науч.-практич. конф., 23 – 24 апреля 2015 г.: тезисы докл. – Д.: ДНУЖТ, – 2015. – с. 168
12. Шинкаренко В. И., Куропятник Е. С. Формализованная спецификация текста и его графовой модели средствами конструктивно-продукционных структур// Проблеми математичного моделювання: Всеукраїнська наук.-метод. конф., 27 – 29 травня 2015 р.: тези доп. – Д.: Видавець Біла О.К., – 2015. – с. 109 – 112
13. Куропятник Е. С. Формализация процессов формирования языковых конструкций как результата мыслительной деятельности человека// Компьютерное моделирование и оптимизация сложных систем: I Всеукраинская науч.-практич. конф., 03 – 05 ноября 2015 г.: тезисы докл. – Д.: ГВУЗ УГХТУ, – 2015. – с. 55 – 57
14. Куропятник Е. С. Оценивание степени уникальности текстов с учетом их злоумышленных изменений // Современные информационные и коммуникационные технологии на транспорте, в промышленности и образовании: IX междунар. науч.-практич. конф., 16 – 17 декабря 2015 г.: тезисы докл. – Д.: ДИИТ, – 2015. – с. 111
15. Шинкаренко В. И. Использование конструктивно-продукционных структур в задачах формализации естественного языка / В. И. Шинкаренко, Е. С. Куропятник // Теоретичні та прикладні аспекти побудови програмних систем. TAAPSD’2016. Тези доповідей. Київ, 2016. – С. 280-284.
16. Куроп’ятник О. С., Шинкаренко В. І. Інструментальне програмне забезпечення для виявлення запозичень у текстах // Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті: X міжнарод. наук.-практич. конф., 14 – 15 грудня 2016 р.: тези доп. – Д.: ДІІТ, – 2016. – с. 147-148
17. Шинкаренко В. І. Конструктивно-продукційна модель графу керування програми/ В. І. Шинкаренко, О. С. Куроп’ятник // Інформаційні технології в моделюванні: Матеріали II-ої всеукраїнської наук.-практич. конф. студентів, аспірантів та молодих вчених (23-24 березня 2017 р., м. Миколаїв). – Миколаїв: МНУ імені В.О. Сухомлинського, – 2017. – С. 40–41
18. Шинкаренко В. І. Використання конструктивно-продукційних структур для обробки мовних конструкцій / В. І. Шинкаренко, О. С. Куроп’ятник // Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті: XI міжнарод. наук.-практич. конф., 13 – 14 грудня 2017 р.: тези доп. – Д.: ДІІТ, – 2017. – с. 154
19. Куроп’ятник О. С. Моделювання текстових модифікацій в задачах виявлення запозичень / О. С. Куроп’ятник // Інформаційні технології в металургії та машинобудуванні. ІТММ’2018: тези доповідей Десятої міжнародної науково-практичної конференції (Дніпро, 27 – 29 березня 2018 р.) /Міністерство освіти

і науки України, Національна металургійна академія України, Дніпропетровський національний університет імені О. Гончара, Дніпропетровський національний університет залізничного транспорту імені академіка В. Лазаряна та ін. – Дніпро: НМетАУ, 2018. – с 164.

20. Куроп'ятник О. С. Образна графова модель мовних конструкцій для виявлення семантичних запозичень / О. С. Куроп'ятник, В. І. Шинкаренко, // Проблеми математичного моделювання: матеріали Всеукр. наук.-метод. конф., 23-25 трав. 2018 р. – м. Кам'янське: ДДТУ, 2018. – 293.
21. Куроп'ятник О. С. Система виявлення запозичень з фільтрацією результатів / О. С. Куроп'ятник // Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті: XII міжнарод. наук.-практич. конф., 12 – 13 грудня 2018 р.: тези доп. – Д.: ДІІТ, – 2018. – с. 94
22. Куроп'ятник О. С. Моделювання і програмна реалізація графового представлення текстів для виявлення запозичень / О. С. Куроп'ятник // Інформаційні технології в металургії та машинобудуванні. ІТММ'2019: тези доповідей міжнародної науково-практичної конференції (Дніпро, 26 – 28 березня 2019 р.) / Міністерство освіти і науки України, Національна металургійна академія України, Дніпропетровський національний університет імені О. Гончара, Дніпропетровський національний університет залізничного транспорту імені академіка В. Лазаряна та ін. – Дніпро: НМетАУ, 2019. – с. 156

Авторські свідоцтва

23. Шинкаренко В. І., Куроп'ятник О.С. Комп'ютерна програма «Система контролю плагіату в студентських роботах» / Свідоцтво про реєстрацію авторського права на твір № 68137 від 05.10.2016.
24. Куроп'ятник О. С., Шинкаренко В. І. Комп'ютерна програма «Система формування тестів для програм-антиплагіатів» («ChangeText») / Свідоцтво про реєстрацію авторського права на твір № 87131 від 22.03.2019.

АНОТАЦІЯ

Куроп'ятник О. С. Конструктивно-продукційні моделі природомовних текстів для виявлення запозичень у структурованих документах – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи. – Національна металургійна академія України, Дніпро, 2020.

Дисертаційну роботу присвячено вирішенню актуальної науково-прикладної задачі розробки моделей природомовних текстів для виявлення запозичень у структурованих документах

На основі розроблених конструктивно-продукційних моделей мови, мовних конструкцій (текстів) та їх графового представлення запропоновано метод і алгоритми зіставлення текстів та структурованих документів для виявлення запозичень.

Запропонована модель процесу маскування запозичень дозволила формалізувати сценарії маскування, створити платформу для моделювання нових змін тексту, автоматизувати побудову тестів для систем антиплагіату.

Розроблено програмні засоби виявлення запозичень у текстових фрагментах та структурованих документах. Розроблено програмний засіб автоматизованого формування тестів для перевірки здатності демаскування запозичень систем анти плагіату.

Комплексне використання отриманих у роботі результатів дозволяє виконувати автоматизовану перевірку текстових фрагментів і структурованих документів на наявність запозичень; тестувати системи антиплагіату та постійно збільшувати тестову базу, будуючи нові сценарії маскувань.

Ключові слова: конструктивно-продукційне моделювання, природномовні тексти, мовні конструкції, виявлення запозичень, маскування запозичень, об'єктно-орієнтоване моделювання, образи, конструктори, стиснення графа, порівняння структурованих документів

АННОТАЦИЯ

Куропятник Е. С. Конструктивно-продукционные модели естественно-языковых текстов для обнаружения заимствований в структурированных документах - На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 01.05.02 – математическое моделирование и вычислительные методы. – Национальная металлургическая академия Украины, Днепр, 2020.

Диссертационная работа посвящена решению актуальной научно-прикладной задачи разработки моделей естественноязыковых текстов для обнаружения заимствований в структурированных документах

На основе разработанных конструктивно-продукционных моделей языка, языковых конструкций (текстов) и их графового представления предложен метод и алгоритмы сопоставления текстов и структурированных документов для выявления заимствований.

Предложенная модель процесса маскировки заимствований позволила формализовать сценарии маскировки, создать платформу для моделирования новых изменений текста, автоматизировать построение тестов для систем антиплагиата.

Разработаны программные средства обнаружения заимствований в текстовых фрагментах и структурированных документах. Разработано программное средство автоматизированного формирования тестов для проверки способности демаскировка заимствований системами антиплагиата.

Комплексное использование полученных в работе результатов позволяет выполнять автоматизированную проверку текстовых фрагментов и структурированных документов на наличие заимствований; тестировать системы антиплагиата и постоянно увеличивать тестовую базу, строя новые сценарии маскировок.

Ключевые слова: конструктивно-продукционное моделирование, естственноязыковые тексты, языковые конструкции, выявление заимствований, маскировка заимствований, объектно-ориентированное моделирование, образы, конструкторы, сжатие графа, сравнение структурированных документов.

ABSTRACT

Kuropiatnyk O. Constructive- synthesizing models of natural language texts for text borrowings detection in structured documents. - Manuscript.

Thesis for obtaining the candidate degree (Ph.D) in engineering sciences in the specialty 01.05.02 – Mathematical modeling and computational methods (Technical science). – The National Metallurgical Academy of Ukraine, Dnipro, 2020.

The dissertation is devoted to solving the relevant scientific applied problem of development natural language texts models for detecting borrowings in structured documents.

The dissertation reviews and analyzes existing models of natural languages, methods for processing language constructions (texts) and comparing natural language texts. Most models do not take into account the performer model and its features, and ignore human thinking processes. Most methods for preprocessing text do not allow restoring the input data after working with them. Methods based on simple comparisons (fingerprints, greedy comparisons, etc.) are sensitive to borrowings disguise mechanisms. Existing software, including those developed using these methods and algorithms, does not take into account the structure of the document to which the text belongs, when checking for borrowings.

According to the results of the analysis of the current state and tendencies of development of linguistic constructions formalization methods and constructions comparison, the necessity of developing effective models and methods of natural language texts for borrowings detection in structured documents is shown.

Constructive-synthesizing modeling (CSM) based on the use of formal languages and grammar apparatus, graph theory, methods and means of set theory, regression analysis were used to solve tasks of developed models of natural language texts and methods for its processing. In the framework of the CSM constructors and methods of their transformation (specialization, concretization, interpretation and realization) are used.

Formalization of the forming human images processes by means of object-oriented modeling was performed. That allowed constructing a hierarchy of images based on the commonality of attributes in order to represent the meaning of the word and to reflect its connection with the objects of reality within the concept of word semantics. It was used in constructing the constructive-synthesizing languages model.

Constructive-synthesizing and object-oriented models of the natural language and text, structured document model, process model of disguise of borrowings text were developed.

Based on the developed models of the language, language constructions (texts) and their graph representation, method and algorithms are proposed for compare text fragments and structured documents to borrowings detection. Computer implementations of models the text graph representation and processes of disguise was created.

The text-weighted graph compression method was developed to improve the performance of the computer implementation of the graph representation model, which made it possible to use the object serialization mechanism to form a structured documents database.

These implementations are software for detecting borrowings in text fragments and structured documents and automated test generation to test the ability to unmask borrowings of anti-plagiarism systems.

The developed models and tools form a single complex, which covers: language and speech – the creation of language constructions; lexical, syntactic, semantic components of language constructions; processes of disguise and detection of borrowings text.

The time effectiveness of the implementation of models the graphical representation text is investigated. The check time for one document was from 11 to 65 sec for the database from 0.6 to 3.8 million characters. Restore graphs spent about 94% of the time. The influence of masking borrowings on increasing the originality of documents amounted to about 0.007%.

Functional effectiveness metrics of developed software for borrowing in text-unstructured documents was compared to its counterpart (WCopyfind). The difference does not exceed 5%. The factors that cause the difference in the performance of the programs have been identified.

The proposed model of the borrowing disguise process allows formalizing masking scenarios, creating a platform for modeling new text changes, and automating the construction of tests for anti-plagiarism systems.

The integrated use obtained results allows performing an automated borrowings check of text fragments and structured documents; performing test anti-plagiarism systems and constantly increasing the test base by building new disguise scenarios.

The developed software allows completing the requirement of the Law of Ukraine "On Higher Education" regarding the academic plagiarism detection in the diploma works of students in the specialty 121 "Software Engineering" at Dnipro National University of Railway Transport named after Acad. V. Lazaryan.

Keywords: constructive-synthesizing modeling, natural language texts, language constructions, borrowings texts detection, borrowings texts disguise, object-oriented modeling, images, constructor, compression of graph, structured documents comparison.

Підписано до друку 06.02.2020. Формат паперу $60\times84^{1}/_{16}$
Ум. друк. арк. 0,9. Обл.-вид. арк. 1,0. Наклад 100 пр. Зам. № 32.

Видавництво ПФ «Стандарт-Сервіс»
Свідоцтво ДК № 3197 від 28.05.2008 р.
52005, Україна, Дніпропетровська обл., смт Слобожанське,
вул. Василя Сухомлинського, 68, кв. 65