Міністерство освіти і науки України
Національний університет «Львівська політехніка»

Швороб Ірина Богданівна

УДК 004.652.4+004.827

МЕТОДИ ТА ЗАСОБИ ЕКСТРАКЦІЇ ТА АНАЛІЗУ СЛАБОСТРУКТУРОВАНИХ ТЕКСТОВИХ ДАНИХ НА ОСНОВІ ДОКУМЕНТО-ОРІЄНТОВАНОГО ГРАФА

10.02.21 – структурна, прикладна та математична лінгвістика

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Львів – 2018

Дисертацією є рукопис.

Робота виконана в Національному університеті «Львівська політехніка» Міністерства освіти і науки України.

Науковий керівник доктор технічних наук, професор
Шаховська Наталія Богданівна,
Національний університет «Львівська політехніка»,
завідувач кафедри систем штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Лупенко Сергій Анатолійович,
Тернопільський національний технічний університету імені Івана Пулюя,
професор кафедри комп'ютерних систем та мереж,

кандидат технічних наук
Надутенко Максим Вікторович,
Український мовно-інформаційний фонд НАН України, завідувач відділу інформаційних технологій

Захист відбудеться 15 березня 2018р. о 14 годині на засіданні спеціалізованої вченої ради Д 35.052.05 у Національному університеті «Львівська політехніка» (79013, м. Львів, вул. С.Бандери, 12, ауд. 226 головного корпусу).

З дисертацією можна ознайомитись у науково-технічній бібліотеці Національного університету «Львівська політехніка» (79013, м. Львів, вул. Професорська, 1).

Автореферат розісланий «14» лютого 2018 р.

Учений секретар
спеціалізованої вченої ради,
доктор технічних наук, професор Р.А.Бунь

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Дисертаційну роботу присвячено розробленню технологій екстракції, структурування, збереження та аналізу слабоструктурованих природномовних текстів.

Актуальність теми. Швидке збільшення кількості інформації зумовило пошук нових підходів до вирішення проблеми з її опрацюванням та збереження. За дослідженнями IDC Digital Universe Study понад 80% світових даних – неструктуровані або слабоструктуровані. Прикладом таких даних є статті, медичні довідки, дописи в соціальних мережах тощо. Наявність таких великих обсягів інформації спонукає до розроблення нових методів та засобів її аналізу. Значний внесок в опрацювання слабоструктурованої інформації українською мовою зробили В.А.Широков, С.А.Лупенко, І.Г.Данилюк, В.В.Балабін, В.В.Робейко та ін. Структурування тексту за допомогою продукційних правил здійснено F.Ciravegna, автоматична розмітка тексту описана в роботах M. Lisboa, автоматичним формуванням списку маркерів займались J.Andersen та M.Puppe. Проте, існуючі на сьогодні методи опрацювання слабоструктурованої інформації суттєво залежать від встановленої розмітки та існуючого корпусу мови. Таким чином, аналіз слабострутурованих даних виконується напівавтоматично, оскільки для текстів з невідомою структурою необхідно визначати розмітку. Тому розроблення методів аналізу природномовних текстів різними мовами є актуальним завданням.

Існує низка недостатньо опрацьованих наукових завдань, що стоять на заваді ефективній організації роботи зі слабоструктурованими природномовними текстами, а саме:

– існуючі підходи до опрацювання неструктурованих та напівструктурованих даних на сьогоднішній день мають емпіричний підхід;

– відсутність систематизації та теоретичного обґрунтування використовуваних методів та засобів опрацювання таких даних негативно впливає на застосування нових методик та методів;

– існуючі методи опрацювання слабоструктурованих даних прив’язані до семантики даних;

– на теперішній час існує незначна кількість україномовних аналізаторів слабоструктурованих текстових даних.

Також важливим завданням є подальше опрацювання витягнених даних з тексту. Зазвичай існуючі на сьогодні засоби зберігають мовні одиниці у реляційних базах даних. Це спричиняє часткову втрату зв’язків, які існували в тексті до екстракції, а отже, погіршує якість опрацювання таких даних.

Отже, наукове завдання розроблення технологій екстракції, структурування, збереження та аналізу слабоструктурованих природномовних текстів є актуальним.

Зв'язок роботи з науковими програмами, планами, темами. Дисертація виконана в межах науково-дослідних робіт «ДБ/ Інтелектуальні інформаційні технології багаторівневого управління енергоефективністю регіону», № держреєстрації 0117U004450 (автор розробила структуру документо-орієнтованої графової бази даних та метод екстракції слабоструктурованих даних) та «Комплекс інтелектуальних інформаційних технологій інтеграції даних для обліку та аналізу підвищення кваліфікації вчителів» № держреєстрації 0113U005273 (автор розробила модель для зберігання слабоструктурованих даних).

Мета і завдання дослідження. Метою дисертаційної роботи є розроблення є методів та засобів екстракції, структурування, збереження слабоструктурованих текстових даних для їх подальшого аналізу.

Мета дисертаційної роботи визначає необхідність розв'язання таких завдань.

1. Аналіз та узагальнення методів опрацювання слабоструктурованих даних.
2. Розроблення інформаційної моделі подання слабоструктурованих даних на основі документо-орієнтованого графа.
3. Розроблення методу опрацювання слабоструктурованих даних на основі документо-орієнтованого графа.
4. Розроблення прикладного програмного забезпечення для опрацювання слабоструктурованих даних та апробація роботи.

Об'єкт дослідження – процес опрацювання слабоструктурованих природномовних текстів.

Предмет дослідження – методи та засоби екстракції, збереження, аналізу та структурування природномовних текстів.

Методи дослідження. Дослідження, виконані під час роботи над дисертацією, ґрунтуються на основі використання теорії графів, статистичного аналізу, методів об'єктно-орієнтованого проектування.

Наукова новизна одержаних результатів. Отримано такі нові наукові результати:

– *вперше* слабоструктуровані природномовні тексти подано у вигляді документо-орієнтованого графа, що дало змогу використати теорію графів для встановлення зв'язків між елементами документа та знизити надлишковість результатів пошуку за запитом користувача;

– *удосконалено* метод екстракції даних з текстових блоків за прагматичною ознакою з метою формування документо-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень;

– *набув подальшого розвитку* метод первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його

подальшого аналізу та опрацювання на основі поділу тексту на блоки за прагматичними ознаками.

Практичне значення одержаних результатів полягає у наступному:

– розроблено алгоритм первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його подальшого опрацювання;

– розроблено алгоритм перерахунку ваг ребер документо-орієнтованого графа, що дає можливість відсіювати надлишковість даних при запиті;

– удосконалено алгоритм екстракції даних з текстових блоків шляхом формування документ-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень та на 8 % збільшити кількість збережених структурних одиниць;

– на основі розробленої архітектури побудовано та впроваджено мовно-інформаційну систему екстракції слабоструктурованих природномовних текстів та роботи з ними.

Одержані в дисертаційній роботі результати використано під час розроблення прототипу системи та впроваджено у ІІ травматологічному відділенні КМКЛШМД м.Львова (аналіз інструкцій для медичних препаратів) та у ТЗоВ «To-You Sp. Z o.o.» (аналіз резюме).

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. У друкованих працях, опублікованих у співавторстві, внесок здобувача такий: [2,4] – розроблено алгоритм пошуку нечітких дублікатів; [6,8] – введено поняття текстового шаблону та маркерів як елементів методу екстракції даних з природномовних текстів.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на семінарах та конференціях:

− X Міжнародна науково-практична конференція «Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті» (Дніпро, 2016);

− Міжнародна конференція «The experience of designing and application of CAD systems in microelectronics» CADSM (Львів-Поляна, 2015);

− X Міжнародна конференція «Комп'ютерні науки та інформаційні технології» CSIT (Львів, 2015);

− XVII Міжнародна науково-технічна конференція «Штучний інтелект і інтелектуальні системи» (Київ, 2017).

Публікації. Основні результати дисертаційної роботи висвітлено в 8 друкованих працях, у тому числі трьох статтях в наукових періодичних виданнях інших держав [1, 2, 4], 3 – у наукових фахових виданнях України [3, 5, 6], 2 – у матеріалах конференцій [7-8].

Структура роботи. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел та додатку. Загальний обсяг дисертації 176 сторінок, з яких основного тексту – 130 сторінок. Робота містить 9 таблиць та 55 рисунків. Список літератури містить 104 найменування.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** наведено загальну характеристику роботи, обґрунтовано актуальність теми, визначено об'єкт і предмет дослідження, сформульовано мету і завдання дослідження, розкрито застосовані методи дослідження, наукову новизну, практичне значення і апробацію одержаних результатів.

У **першому розділі – «Аналітичний огляд технологій роботи зі слабоструктурованими даними»** – систематизовано джерела та з'ясований стан наукових досліджень за темою дисертації. Зокрема, проведено аналіз існуючих засобів екстракції даних, досліджено методи синтаксичного розбору тексту та здійснено їх аналіз, розглянуто поняття слабоструктурованих даних та здійснено огляд основних способів їх представлення. Також у роботі здійснено аналіз існуючих методів пошуку нечітких дублікатів у текстах.

Слабоструктуровані дані мають певні особливості:

• структура даних може бути неповною, частково визначеною, а також допускати виключення;

• значення скалярних даних представлені у вигляді текстової інформації;

• виникає проблема визначення приналежності даних, тому що не завжди можна однозначно визначити коректність оброблюваного документа.

Основні проблеми в роботі зі слабоструктурованими даними: різноманітність даних; можливість розширення; зберігання.

Модель слабоструктурованих даних повинна враховувати вище перераховані особливості. Виділено основні проблеми, що виникають під час розроблення моделі слабоструктурованих даних:

1) у процесі роботи з даними заздалегідь невідомий ступінь їх коректності і, як наслідок, у моделі необхідний інструментарій для оцінки «правильності» даних; враховуючи, що в слабоструктурованих даних усі атрибути представлені у вигляді текстової інформації, необхідний досить гнучкий механізм перевірки приналежності даних до конкретного атрибута;

2) схема даних може або зовсім не існувати, або не повною мірою відповідати оброблюваним даним; оскільки працювати з документом, не маючи ніяких уявлень про його структуру, неможливо, виникає завдання виділення схеми з оброблюваних даних, а також її коректування у процесі експлуатації моделі й одержання нової інформації;

3) деякі атрибути даних можуть бути або взагалі відсутні, або не повною мірою задовольняти умовам коректності, заданим для цих атрибутів; таким чином, у цій моделі повинен існувати інструмент обробки виключень, що дає змогу формувати спосіб запиту до цих даних, ґрунтуючись на заздалегідь заданих критеріях.

Здійснено аналіз розглянутих способів представлення слабоструктурованих даних на основі основних моделей: наявність екземпляра документа, наявність схеми документа, наявність ідентифікаторів наборів елементів, наявність бінарних та *n*-арних наборів зв'язків, обмеження участі наборів елементів у наборах відношень, наявність посилань між наборами елементів, наявності атрибутів і наборів елементів, наявності атрибутів множин зв'язків, ступеню впорядкування наборів елементів і атрибутів. Результати порівняння подану у табл. 1.

Таблиця 1. Порівняння представлень слабоструктурованих даних

	DTD	DOM	OEM	S3-graph	CM	EER	XML Tree	ORA-SS
Екземпляр	-	+	+	-	-	-	+	+
Схема	+	-	+	+	+	+	+	+
Виявлення атрибуту	+ -	-	-	-	-	+ -	-	+
Обмеження участі	+ -	-	-	+ -	+ -	+	+ -	+
Посилання	+ -	-	-	+ -	-	+	+ -	-
Атрибути та елементи	+ -	+	-	+	-	+	+	+
Атрибути відношень	-	-	-	-	+ -	+	-	+
Упорядкування	+ -	-	-	-	-	+	+ -	+

Синтаксичний аналіз (парсинг) є важливою складовою опрацювання тексту і спрямований на розпізнавання, виділення та групування даних. Використовуючи синтаксичний аналіз можна дуже швидко опрацьовувати великі об'єми інформації, оскільки вручну це робити практично неможливо. Загалом парсинг є ефективним рішенням для автоматизації збору та зміни інформації.

Будь-який аналізатор складається з трьох частин, які відповідають за три окремі процеси розбору.

• Отримання тексту у його первісному вигляді. Отримання тексту часто означає завантаження текстового документа, з якого необхідно отримати певні дані.

• Вилучення та перетворення даних. Необхідні дані, які були отримані на першому етапі, на цій фазі «витягаються». Для вилучення найчастіше

використовуються регулярні вирази. Крім того, на цьому етапі здійснюється перетворення витягнутих даних у певний формат, якщо це необхідно.

• Генерування результатів. Це заключний етап аналізу. Він досягається шляхом виводу або запису даних, отриманих на другому етапі, у бажаному форматі. Часто запис здійснюється безпосередньо в базі даних.

Завдання синтаксичного аналізатора полягає у визначенні, чи є і в який спосіб вхідні дані можуть бути отримані з початкового символу граматики.

У роботі здійснено аналіз існуючих синтаксичних аналізаторів, таких як Earley парсер, LL парсер, метод рекурсивного спуску, CKY парсер, LALR, парсер Prett парсер за такими параметрами: підтримка граматики, метод синтаксичного розбору, напрям проходу, наявність повторень. В табл. 2 подано узагальнену характеристику проаналізованих алгоритмів.

Таблиця 2. Характеристика алгоритмів синтаксичних аналізаторів

	Earley парсер	LL парсер	Метод рекурсивного спуску	CKY парсер	LALR парсер	Prett парсер
Підтримка граматики	Всі	Всі	Всі	CNF	Всі	—
Метод синтаксичного розбору	Нисхідний	Нисхідний	Нисхідний	Висхідний	Висхідний	Нисхідний
Напрям проходу	Зліва направо	Зліва направо	Зліва направо	Справа наліво	Справа наліво	Зліва направо
Наявність повторень	Є	Є	Є	Є	—	Є

Алгоритми порівнювались за двома критеріями: швидкодією опрацювання граматики та швидкодією опрацювання поданого на вхід тексту. Порівняння характеристик виконання вищеописаних алгоритмів представлено в табл. 3.

Таблиця 3. Результати порівняння

	Earley парсер	LL парсер	Метод рекурсивного спуску	CKY парсер	LALR парсер	Prett парсер
Опрацювання граматики, мс	59	4	36	20	19	46
Опрацювання тексту, мс	59	4	35	20	19	44

У роботі не виконано повний граматичний аналіз, але показано, що тексти з відомою структурою аналізуються швидше. Набагато легше реалізувати синтаксичний аналіз в слабоструктурованих текстах, які розділені на блоки. Вони можуть бути організовані за допомогою особливих характеристик даних (шаблонів), що створюються з використанням вже

отриманої інформації. Такі характеристики витягуються з вхідного документа і використовуються для ідентифікації інформації. Цей підхід може бути використаний для всіх видів інформації.

Варто відзначити, що деякі алгоритми працюють швидше, ніж інші. Але в це й же час не всі алгоритми можуть працювати з усіма граматиками. Наприклад, CKY парсер швидший, ніж алгоритм Earley, але для CKY потрібно, щоб граматика бути в CNF, а алгоритм Earley працює для будь-якої граматики. Вирішення цієї проблеми може полягати в удосконаленні існуючих алгоритмів або створенні нових алгоритмів шляхом їх комбінацій.

Аналіз методів та засобів опрацювання текстової інформації показав, що на теперішній час залишилися нерозв'язаними такі задачі:

- існуючі підходи до опрацювання неструктурованих та слабоструктурованих даних недостатньо точно виділяють конкретні текстові об'єкти із природномовних текстів;
- існуючі способи представлення слабоструктурованих даних не підтримують всіх основних вимог для побудови моделей даних.
- велика кількість даних та її швидке збільшення зумовлює пошук нових підходів до її оптимального збереження.

Результати розділу опубліковано у [5].

У **другому розділі – «Розроблення моделі подання слабоструктурованих даних»** – здійснено формальну постановку задачі розроблення моделі подання слабоструктурованих даних на основі документо-орієнтованого графа. Уведено перелік концептуальних понять та визначень. Здійснено порівняльну характеристику запропонованої документо-орієнтованої графової бази даних з існуючими графовою та документо-орієнтованою базами даних. Використано елементи теорії графів при роботі зі слабоструктурованими графами. Розроблено метод перерахунку ваг ребер документо-орієнтованого графа.

Уведено перелік основних понять та означень:

Слабоструктуровані дані T – дані, для яких визначені деякі правила і формати, але в найзагальнішому вигляді. Вони не організовані спеціальним, наперед заданим, чином, що робить доступ і можливість аналізу складним завданням. Проте, вони можуть мати інформацію, пов'язану з ними, наприклад теги метаданих, що дає змогу отримати доступ до таких даних.

Прагматична ознака Pr – назва концептів ключових ознак тексту, за якими будується структура текстового документа, *Pr*=(*Name*, *Rank*), де *Name* – назва прагматичної ознаки, *Rank* – важливість прагматичної ознаки для подальшого аналізу тексту.

Текстовий шаблон Ts – організований та впорядкований за множиною прагматичих ознак тексту *Pr* у відповідності з певною цільовою установкою експертів, характеристик тексту на основі форматування та метаданих, що

дозволяє ділити слабоструктуровані тексти на блоки, придатні для подальшого опрацювання та поділу на структурні одиниці.

Структурна одиниця SE– слово або словосполучення, що містить певне смислове навантаження та відіграє роль назви структури даних у слабоструктурованому тексті.

Екстракція даних – процес видобування структурних одиниць з слабоструктурованого природномовного тексту, який складається з таких етапів:

$$f_1: T \to \{Pr\}, \tag{1}$$

$$f_2: Ts \to \{SE\}, Ts \in Pr_i, Rank_i = \frac{N^{Rank_i}}{\sum_{i=1}^{k} N^{Rank_i}}, Rank_i > e, \tag{2}$$

де f_1 – функція поділу слабоструктурованого тексту T на текстові шаблони Ts за прагматичними ознаками Pr, f_2 – функція формування множини структурних одиниць SE за прагматичними ознаками, важливість прагматичної ознаки визначається як середньозважена частота зустрічі назви прагматичної ознаки у запитах користувачів. Аналізуватимемо тільки ті прагматичні ознаки, значення важливості яких вище за встановлений експертом поріг e.

Для зберігання структурних одиниць та пошуку інформації на їх основі у роботі здійснено аналіз існуючих нереляційних баз даних, а саме баз даних типу «ключ-значення», стовпчикових, документо-орієнтованих, графових та об’єктно-орієнтованих баз даних. Для детального розгляду та аналізу обрано графову та документо-орієнтовану бази даних. Також уведено модель документо-орієнтованого графа.

Під час роботи зі слабоструктурованими даними важливо зберегти якомога більшу їх кількість у найшвидшій для використання формі. База даних на основі документо-орієнтованого графа включає складність вузла графа даних, тобто, коли вузлом є елемент з багатьма різними характеристиками. Така реалізація використовує документ для забезпечення гнучкості запитів до графових баз даних, а збереження у вигляді «ключ-значення» забезпечує швидкий пошук даних.

Об'єкт G такої бази даних подано так:

$$G = (N, E),$$

де N – множина вершин графа, N=$\{n_1, \dots n_m\}$, E – множина ребер, $E = \{e_1, \dots, e_n\}$.

З огляду на те, що документ використовує спрямований граф, то вузол графа може бути представлений у вигляді об'єкта, який містить множину параметрів типу ключ-значення $\langle k, v\rangle$, а також значення типу вершини $NodeType$, наприклад,

$$N = \{\langle k, v\rangle, NodeType\}, v_i \in \{SE\}.$$

Ребро графа представлено у вигляді об’єкта, який містить вказівники на батьківську $ParentNode$ та дочірню $ChildNode$ вершини, значення типу ребра $EdgeType$, $EdgeType \in Pr$ та вагу $Weight$:

$$e_i = \{ParentNode_i, ChildNode_i, EdgeType_i, Weight_i\}.$$

Операції над графами подано таким чином:

1) Об’єднання графів $G_1(N_1,E_1)$ та $G_2(N_2,E_2)$ з множиною вершин, що перетинаються:

a. додавання вершин, що не перетинаються:

$$N_{1k} = N_1 \backslash N_2, N_{2k} = N_2 \backslash N_1,$$
$$N_k = N_{1k} \cup N_{2k}, E_k = E_{1k} \cup E_{2k},$$

b. додавання спільних вершин та перерахунок ваг дуг, що з цих вершин виходять або входять:

$$\forall e_1 \notin E_k, \forall e_2 \notin E_k : e(Weight) = e_1(Weight_i) + e_2\big(Weight_j\big):$$
$$e_1(ParentNode_i) \in N_2,$$
$$e_2(ParentNode_j) \in N_1, e_1(ChildNode_i) \in N_2,$$
$$e_2(ChildNode_j) \in N_1, e_1(EdgeType_i) = e_2(EdgeType_j).$$
$$E_k = E_k \cup \{e\}.$$
$$N_k = N_k \cup \{ParentNode_i, ChildNode_j\}.$$

2) Перетин графів $G_1(N_1,E_1)$ та $G_2(N_2,E_2)$:

$$N_k = N_1 \cap N_2, E_k = E_1 \cap E_2.$$

3) Паралельне з'єднання графів $G_1(N_1,E_1)$ та $G_2(N_2,E_2)$:

a. Пошук термінальної пари – стоку S та витоку T:

$$S: G_1(e(ParentNode_i)) = G_2\left(e\big(ParentNode_j\big)\right), G_1(e(EdgeType_i)) =$$
$$= G_2\left(e\big(EdgeType_j\big)\right);$$
$$T: G_1(e(ChildNode_i)) = G_2\left(e\big(ChildNode_j\big)\right), G_1(e(EdgeType_i)) =$$
$$= G_2\left(e\big(EdgeType_j\big)\right);$$

b. Об’єднання стоків з G_1 та G_2;

c. Об’єднання витоків з G_1 та G_2;

d. Об’єднання G_1 G_2;

4) Послідовне з'єднання графів $G_1(N_1,E_1)$ та $G_2(N_2,E_2)$:

a. Пошук термінальної пари – стоку S та витоку T, так, як і для паралельного з’єднання;

b. Об’єднання витоку з G_1 з стоком G_2;

c. Об’єднання G_1 G_2.

Розглянемо приклад збереження і обробки даних про лікарські засоби, побудувавши документ-орієнтований граф на основі бази даних Neo4j.

Створені вершини графа будуть двох типів: $NodeType = \{\text{Препарат}, \text{Хвороба}\}$. Ребра також будуть двох типів:

$EdgeType = \{\text{Показання}, \text{Протипоказання}\}$. Рис. 1демонструє створений документо-орієнтований граф для медичних препаратів.

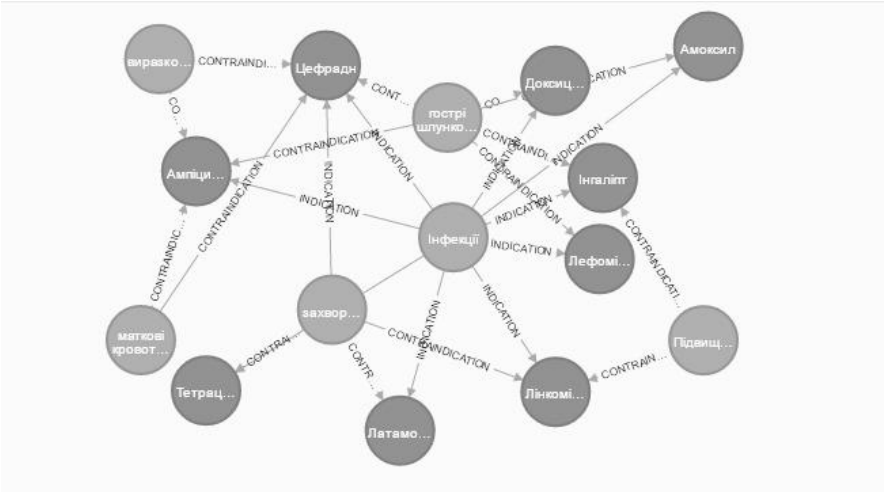

Рис. 1. Документо-орієнтований граф для системи збереження та обробки інформації про медичні препарати

Подальше узагальнення відображення зв'язків між об'єктами слабоструктурованих даних за допомогою документо-орієнтованих графових баз даних полягає у визначенні ребрам та дугам деяких кількісних значень, якісних ознак чи характерних властивостей, які називають вагою.

Далі у роботі розроблено метод перерахунку ваг ребер, що застосовується для видалення надлишкових даних з результату пошуку за вхідними параметрами, за рахунок видалення вершин, вага ребер яких менша за встановлену експертом. Перерахунок ваг ребер визначено так.

Нехай маємо масив *R* попередніх оцінок якості вибору:

$$R = \{\langle x, f\rangle\},$$

де *x* – індивідуальне значення оцінки; f – повторюваність оцінки.

Використовуючи формулу середньозваженого, знаходимо нову вагу між вершиною вхідних даних та вибраним результатом:

$$Weight = \frac{\sum_{i=1}^{n} x_i f_i + r}{\sum_{i=1}^{n} f_i + 1},$$

де *r* – нова введена оцінка.

Далі оновлюємо масив *R*, додавши до нього нове значення оцінки.

Результати розділу опубліковано у [1, 3, 7].

У **третьому розділі – «Розроблення методів екстракції слабоструктурованих природномовних текстів»** – розроблено метод виділення складових елементів для побудови текстового шаблону. Розроблено метод

виділення прагматичних ознак. Побудовано алгоритм пошуку нечітких дублікатів у природномовних текстах. Розроблено методи первинного аналізу тексту та аналізу текстових блоків.

Загалом для екстракції даних зі слабоструктурованих природномовних текстів необхідно звести ці дані до єдиного вигляду, оскільки вони можуть завантажуватись у систему у різних форматах, наприклад html, txt, pdf. Для цього реалізовано видалення форматування, додаткових неінформативних об'єктів (малюнки, підключені скрипти і т.д.), зберігши інформацію про форматування тексту та метадані в окрему базу даних. Ці дані використовуються для перевірки на ідентичність до існуючих текстових шаблонів або побудови нових.

Після перевірки тексту на наявність нечітких дублікатів виконується перевірка тексту на відповідність готовому шаблону, якщо шаблон не знайдено, то здійснюється спроба побудувати шаблон з пошуком прагматичних ознак та з урахуванням форматування та позицій знайдених прагматичних ознак у тексті.

Текстовий шаблон складається з послідовності текстових блоків та утворює кортеж, а текстові блоки – з послідовності слів, яка, у свою чергу, зображується кортежем. Введемо множину прагматичних ознак шаблону, які містяться у досліджуваних текстах. У тексті знаходять прагматичні ознаки.

Текстовий шаблон – це неструктурований або напівструктурований файл, який складається з послідовності речень, а речення – з послідовності слів. З всієї множини слів у документі вибираються тільки ті, що мають змістовне наповнення, тобто формується база даних «Прагматичні ознаки».

Текстовий шаблон, прагматичні ознаки та структурні одиниці подано за допомогою нотації Бекуса-Наура:

<Текст>:= {<Текстовий_блок>}

<Текстовий_блок>:=<Прагматична_ознака> “.” <Список>

<Список>:=<Структурна_одиниця> | [“,”<Список>]

<Прагматична_ознака>:=<Жирний>|<Курсив>|<Підкреслений>|<Відцентро-ваний>|

<Колір>|<Великі букви>

<Структурна_одиниця>:=<Слово>| [“,”<Структурна_одиниця >]

Таким чином, опрацьовуються текстові блоки, на початку яких розміщена прагматична ознака. Необхідно зазначити, що різниця у форматуванні прагматичної ознаки та структурної одиниці має бути визначена і статистично: формат відображення, що використовується для прагматичної ознаки, зустрічається значно рідше, аніж формат структурної одиниці.

Прикладами природномовних текстів, що складаються з текстових блоків, описаних вище, є: інструкції до медичних препаратів, медична картка пацієнта, резюме, технічна документація тощо.

Необхідно зазначити, що для розробленого текстового шаблону неважливою є мова тексту, адже прагматична ознака виділяється за форматуванням.

1) Визначається частота слів з різним типом форматування:
$NB = \frac{count(Bold)}{count(*)}$, $NI = \frac{count(Italic)}{count(*)}$, $NU = \frac{count(Underline)}{count(*)}$.

2) Обираються ті слова (словосполучення), для яких *NB* або *NI* або *NU* мінімальні та не дорівнюють нулеві.

3) Слова з пункту 2 додаються у множину *Pr*.

За знайденим шаблоном відбувається поділ тексту на текстові блоки за значенням прагматичної ознаки *Pr* (функція f_1 з (1)). Готові текстові блоки очищаються від мовних кліше, слів-зв'язок, стоп-слів. Наступним етапом здійснюється поділ текстового блоку на речення, які в свою чергу діляться на словосполучення у залежності від знаків пунктуації (функція f_2 з (2)). У словосполученнях видаляються закінчення та префікси, і виконується перевірка на подібність таких словосполучень до наявних у базі даних, при цьому враховується ранг прагматичної ознаки, оскільки деякі дані несуть більше смислове навантаження існує потреба виділяти їх в якості окремих вершин.

Результати розділу опубліковано у [2, 4, 6,8].

Четвертий розділ – «Розроблення архітектури мовно-інформаційної системи екстракції слабоструктурованих природномовних текстів та роботи з ними» – присвячено комплексній реалізації практичного використання запропонованих підходів. Зокрема у розділі спроектовано архітектуру та розроблено систему екстракції, структурування, збереження та аналізу слабоструктурованих природномовних текстів. Апробовано розроблені методи для роботи зі слабоструктурованими медичними даними. Також розроблені методи використано для формування системи роботи з резюме найманих працівників.

Спроектовано архітектуру мовно-інформаційної системи екстракції слабоструктурованих природномовних текстів та роботи з ними, вона складається з 3-х рівнів (рис. 2): рівень роботи з користувачем; рівень бізнес-логіки, де виконується екстракція даних, їх структурування, збереження та аналіз; рівень даних.

Систему екстракції слабоструктурованих природномовних текстів розділено на модулі та підсистеми. Програма складається з наступних компонент: база даних; підсистема графічного представлення; підсистема опрацювання слабоструктурованих природномовних текстів та запису їх базу даних; підсистема роботи зі структурованими даними у представленні об’єктами документо-орієнтованого графа.

В архітектурі системи передбачено дві бази даних: документо-орієнтована графова база даних, де будуть зберігатись дані, отримані зі

слабоструктурованих текстів, та база даних, що містить загальну інформацію, таку як MD5-суми завантажуваних файлів, словники стоп-слів, ключових слів, частотні словники, словники мовних кліше, метадані, готові текстові шаблони та набори прагматичних ознак (маркерів). Обидві бази даних використовуються іншими модулями розробленої системи.

Підсистема графічного представлення призначена для взаємодії з користувачем та отриманням інформації з бази даних. Є дві частини графічного інтерфейсу системи. Перша частина – графічний інтерфейс системи опрацювання слабоструктурованих природномовних текстів, що призначений для роботи з адміністратором системи. Друга частина – графічний інтерфейс для роботи з користувачами – клієнтська програма системи підтримки прийняття рішень.

Після завантаження файла документу або посилання на web-сторінку з необхідним текстом здійснюється основна логіка програми, тобто використовується підсистема опрацювання слабоструктурованих природномовних текстів та запису їх у базу даних. Модуль аналізу вхідного файлу здійснює перевірку чи опрацьовувався завантажений документ раніше чи ні. У випадку якщо знайдено існуючий у базі даних документ, то користувач одержить відповідне повідомлення. Модуль зведення вхідних даних до єдиного вигляду здійснює первинну обробку вхідних даних та зводить їх до єдиного вигляду, оскільки дані можуть мати різне форматування та метадані. Модуль пошуку нечітких дублікатів здійснює більш детальну перевірку вхідного тексту, з метою уникнення дублювання інформації у базі даних.

Система перевірки та побудови шаблонів призначена для поділу тексту на текстові блоки відносно текстового шаблону та прагматичних ознак. Система аналізу текстових блоків призначена для виділення структурних одиниць з тексту. Система аналізу та збереження результатів аналізує та записує дані в документо-орієнтовану графову базу даних та здійснює зв'язок з користувачем, у випадку коли виникають суперечливі дані та необхідно підтвердити або відхилити запис у базу даних.

Для другої частини користувацького інтерфейсу основну роботу виконує підсистема роботи зі структурованими даними у представленні об’єктами документо-орієнтованого графа. Система опрацювання користувацьких запитів здійснює запити до документо-орієнтованої бази даних та представляє результати користувачеві.

Далі у розділі апробовано розроблені методи перерахунку ваг ребер документо-орієнтованого графа на прикладі роботи системи опрацювання інструкцій до медичних препаратів. Для дослідження обрано 100 пацієнтів з однаковими симптомами: температурою та запаленням.

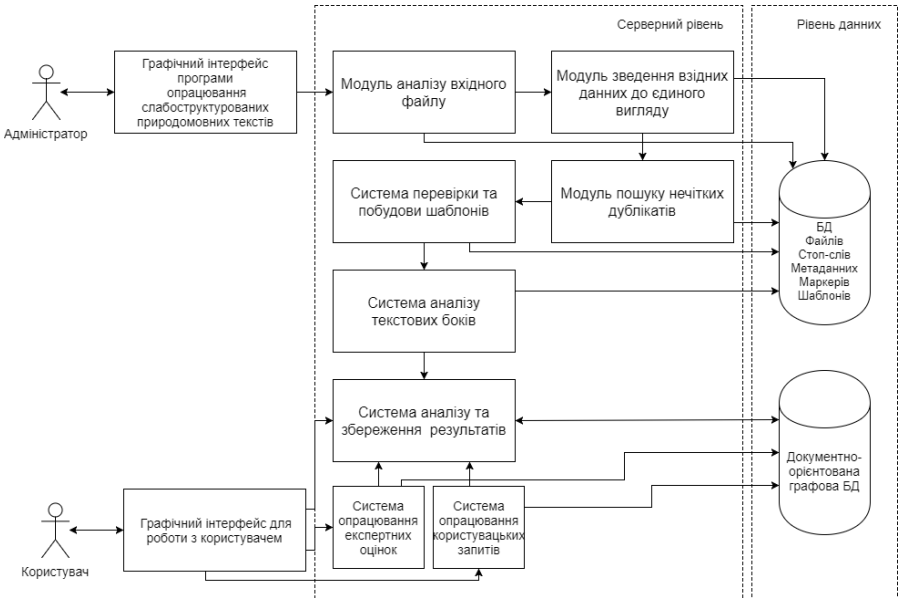

Рис. 2. Архітектура мовно-інформаційної системи екстракції слабоструктурованих природо мовних текстів та роботи з ними

У результаті отримаємо граф, зображений на рис. 3.Створені вершини графа є трьох типів: пацієнт, хвороби та препарати. Ребра будуть двох типів: симптоми та показання. На початку дослідження всі ваги ребер встановлюються в 1. Встановлюємо експертне значення для порівняння з найкоротшим шляхом між вершинами рівним 0,1. Тобто всі шляхи між вершинами симптомів та вершинами препаратів, які менші за 0,1, будуть видалятись.

Нехай для заданих симптомів буде обрано препарат Ампіцилін. Тоді після здійснення вибору отримуємо новий результуючий граф, зображений на рис. 4.

Після проведення лікування, здійснюється повторний огляд пацієнта та лікарем вводиться оцінка якості лікування даним препаратом, тобто визначається чи допоміг обраний препарат вилікувати симптоми пацієнта. Оцінка виставляється в межах від 0 до 1.

Якщо симптом зник повністю, то виставляється оцінка 1, якщо зовсім не зник — то 0. Якщо симптом вилікувано частково, то оцінка виставляється на розсуд лікаря. Якщо симптом вилікувано повністю, то вага ребра між вершинами пацієнта та симптома стає рівною 0, а, отже, вершина даного симптома видаляється з результуючого графа.

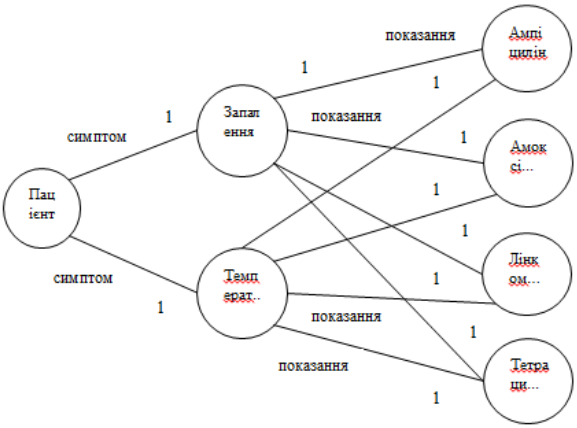

Рис. 3. Результат виконання запиту до документо-орієнтованої графової бази даних

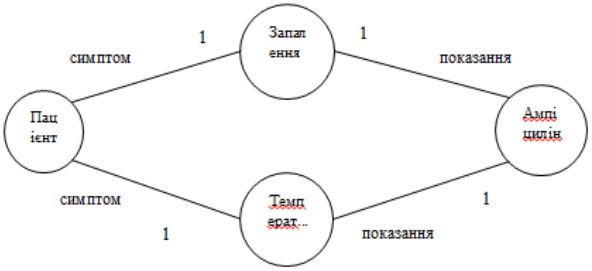

Рис. 4. Граф-відображення призначеного лікування пацієнтові

Після опрацювання вибірки зі 100 пацієнтів з однаковими симптомами, отримуємо граф, зображений на рис. 5.

Після пошуку найкоротшого шляху між вершинами бачимо, що найкоротшим шляхом є відстань між вершинами симптому Запалення та вершиною препарату Лінкоміцин і становить 0,096, що наближається до 0. Отже, вибрану вершину необхідно видалити з графу і при наступному пошуку ця вершина препарату відображатись не буде. Кінцевий граф після проходження запропонованого алгоритму зображено на рис.5.

Отже, запропонований алгоритм є оптимальним рішенням для роботи з великими об’ємами слабоструктурованих даних, оскільки дозволяє відкидати дані з низькою ефективністю при певному вхідному наборі параметрів. Було здійснено аналіз обробки слабоструктурованих даних для різних типів баз даних (табл. 4). Аналіз проводився за такими параметрами: кількість створюваних об'єктів (документів або вузлів) (N), вага бази даних

(*W*), час запису в базу даних (*t*), час виконання запиту з декількома умовами (*tc*). Для аналізу було використано 100 інструкцій для медичних препаратів.

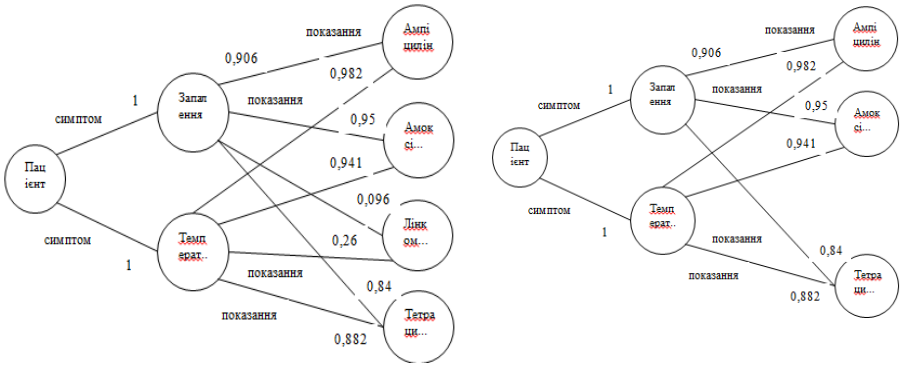

Рис.5. Результат запиту до документо-орієнтованої графової бази даних: а) загальний результат; б) результат після видалення надлишкових даних

Таблиця 4. Результати аналізу обробки слабоструктурованих даних

	N	*W(Мб)*	*t(мс)*	*tc(мс)*
Документо орієнтована БД	100	40.2	10	2
Графова БД	685	30.9	15	1.4
Документо-орієнтований граф	740	61.1	20.72	1.3

Істотною перевагою зберігання даних у документо-орієнтованій базі даних є зручність для подальшої обробки даних. Однак запити будуть складними і при запиті може прийти набагато більше інформації, ніж це необхідно. Це, в свою чергу, впливає на продуктивність. Дані в графовій базі даних займають великий об'єм. Час оптимізації бази даних графа є більшим. Документо-орієнтовані графові бази даних значно домінують над реляційними у виконанні пошуку в реальному часі з великими обсягами даних. Таким чином, графові бази даних доцільно використовувати за наявності великих обсягів даних і ресурсів, за умови, що швидке виконання пошуку є дуже важливим. У залежності від вимог до проекту та до оптимізації і пришвидшення роботи із великими об'ємами даних потрібно використовувати відповідні бази даних. Наприклад, для програмного рішення дуже важливим є швидкий пошук, тому для цього можна використати графову базу даних. В окремих випадках можна зберігати дані у вигляді документо-орієнтованих графових баз даних.

Результати розділу опубліковано у [2, 3,4, 6].

В **додатках** містяться акти впровадження та програмний код.

ОСНОВНІ РЕЗУЛЬТАТИ ТА ВИСНОВКИ

У дисертаційній роботі розв'язано актуальне наукове завдання розроблення технологій для екстракції, збереження, опрацювання та аналізу слабоструктурованих даних. Основні результати дисертаційного дослідження викладені у висновках, які зводяться до наступних положень:

1. Здійснено аналіз моделей слабоструктурованих даних, способів опрацювання природномовних текстів та їх аналізу та пошуку прихованих залежностей даних, що дало змогу здійснити постановку задачі дослідження та виділити раніше нерозв'язані задачі.

2. Уведено поняття зваженого документ-орієнтованого графа для представлення слабоструктурованих природномовних текстів, що дало змогу використати теорію графів для встановлення зв'язків між елементами документа та визначення типу відношення між документом та шаблоном. Описано вершини та ребра кількох типів, що уможливило подання у вигляді графу різних сутностей та властивостей із збереженням їхніх параметрів.

3. Вперше розроблено метод первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його подальшого опрацювання, розділивши його на текстові блоки. Це дало змогу зменшити складність операції пошуку необхідного концепту в природномовному тексті.

4. Удосконалено метод екстракції даних з текстових блоків шляхом формування документ-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень та на 8 % збільшити кількість збережених структурних одиниць.

5. Розроблено систему розуміння природномовних текстів для опрацювання та аналізу даних, архітектура якої відрізняється від існуючих наявністю модулів, які виділяють прагматичні ознаки та розділяють текст на текстові блоки.

6. Впроваджено інформаційно-лінгвістичну систему для аналізу слабоструктурованих текстів у різних предметних областях, зокрема, для автоматичного формування бази даних медичних препаратів та аналізу текстів резюме. Подання відповіді користувачеві у вигляді графа дає змогу не тільки візуалізувати дані, але й показати зв'язки у знайдених даних. За рахунок перерахунку ваг ребер релевантність запиту збільшуватиметься з часом експлуатації системи.

СПИСОК ПУБЛІКАЦІЙ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Shvorob I. B. New approach for saving semistructured medical data / Shvorob I. B. / Advances in Intelligent Systems and Computing. − Vol. 512. – 2017. − P. 29-40. ISSN: 2194-5357. Doi: 10.1007/978-3-319-45991-2_3 (*Scopus*)

2. Shahovska N. B. The method for detecting plagiarism in a collection of documents / Shahovska N. B., Shvorob I. B. / Journal of Applied Computer Science. – Vol. 11, #3. – 2015. – P. 56-66 (*Index Copernicus*).
3. Швороб І. Б. Підхід до роботи зі слабоструктурованими даними на основі використання ваг ребер для документо-орієнтованої бази даних / Швороб І. Б. // Бионика интеллекта. – 2017. – Вип. 1(88). – С. 90-96 (*Index Copernicus*).
4. Shahovska N. B. The intelligent system of determine the degree of resemblance of the texts / Shahovska N. B., Shvorob I. B. / Research Journal of International Studies. – V. 30, Is. 11. − 2014. –P. 87-88 (*РИНЦ*).
5. Швороб І. Б. Порівняльний аналіз методів синтаксичного розбору текстів / Швороб І. Б. // Вісник Національного університету «Львівська політехніка»: Інформаційні системи та мережі. − 2015. – Вип. 814. – С. 197-202 (*Google Scholar*).
6. Шаховська Н. Б. Метод побудови текстового шаблону для екстракції інформації зі слабоструктурованих даних / Шаховська Н. Б., Швороб І. Б. // Штучний інтелект. – 2017. – № 2. – С. 52-61 (*Google Scholar*).
7. Швороб І. Б. Документо-орієнтований граф як засіб збереження слабоструктурованих даних / Швороб І. Б. // Збірник тез X Міжнародної науково-практичної конференції «Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті». − Дніпро : ДНУЗТ, 2016. − С. 68-69.
8. Шаховська Н. Б. Побудова текстового шаблону для екстракції слабоструктурованих даних / Шаховська Н. Б., Швороб І. Б. // Матеріали XVII Міжнародної науково-технічної конференції «Штучний інтелект та інтелектуальні системи» (AIIS'2017) . – Київ, 2017. – С. 235-239.

АНОТАЦІЇ

Швороб І.Б. Методи та засоби екстракції та аналізу слабоструктурованих текстових даних на основі документо-орієнтованого графа. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня кандидата технічних наук за спеціальністю 10.02.21 – Структурна, прикладна і математична лінгвістика. – Національний університет «Львівська політехніка» МОН України, Львів, 2018.

У дисертаційній роботі розв'язано актуальне наукове завдання розроблення технологій для екстракції, збереження, опрацювання та аналізу слабоструктурованих даних.

Здійснено аналіз моделей слабоструктурованих даних, способів опрацювання природномовних текстів, їх отримання, що дало змогу здійснити постановку завдання дослідження. Введено поняття документ-орієнтованого графа для представлення слабоструктурованих природно-

мовних текстів, що дало змогу використати теорію графів для встановлення зв'язків між елементами документа та визначення типу відношення між документом та шаблоном. Вперше розроблено метод первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його подальшого опрацювання.

Удосконалено метод екстракції даних з текстових блоків шляхом формування документ-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень та на 8 % збільшити кількість збережених структурних одиниць. Розроблено систему розуміння природномовних текстів для опрацювання та аналізу даних.

Ключові слова: прагматична ознака, слабоструктуровані дані, документо-орієнтований граф, NoSQL база даних, екстракція даних.

Швороб И.Б. Методы и средства экстракции и анализа слабоструктурированных текстовых данных на основе документо-ориентированного графа. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 10.02.21 – Структурная, прикладная и математическая лингвистика. – Национальный университет «Львивська политэхника» МОН Украины, Львов, 2018.

В диссертационной работе решена актуальная научная задача разработки технологий для экстракции, хранения, обработки и анализа слабоструктурированных данных.

Осуществлен анализ моделей слабоструктурированных данных, способов обработки естественно-языковых текстов, их получения, что позволило осуществить постановку задачи исследования. Введено понятие документ-ориентированного графа для представления слабоструктурированных естественно-языковых текстов, что позволило использовать теорию графов для установления связей между элементами документа и определения типа отношения между документом и шаблоном. Впервые разработан метод первичного анализа данных, который позволяет частично структурировать естественно-языковые тексты для его дальнейшей обработки.

Усовершенствован метод экстракции данных из текстовых блоков путем формирования документ-ориентированного графа, который в отличие от метода на основе использования меры TF-IDF позволяет учесть семантику предложений и на 8% увеличить количество сохранившихся структурных единиц. Разработана система понимания естественно-языковых текстов для обработки и анализа данных.

Ключевые слова: прагматический признак, слабоструктурированные данные, документо-ориентированный граф, NoSQL база данных, экстракция данных.

Shvorob I.B. Methods and means of extraction and analysis of poorly structured text data based on a document-oriented graph. – On the rights of manuscript.

The thesis for the degree of candidate of technical sciences, specialty 10.02.21 – Structural, applied and mathematical linguistics. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2018.

The dissertation solved the problem of developing technologies for extraction, storage, processing and analysis of semistructured data.

The analysis of models of semistructured data, methods of processing natural language texts for their obtaining were carried through, which made it possible to set the task of research. The notion of a document-oriented graph for the presentation of semistructured text-to-speech texts was introduced, which enabled the use of graph theory to establish links between elements of the document and determine the relationship between the document and the template. For the first time, a method of initial analysis of data has been developed, which allows to partially structure the natural language text for its further elaboration. The elements of graph theory are used when working with weakly structured graphs. The method of converting the weights of the edges of a document-oriented graph is developed.

The method of extraction of data from text blocks has been improved by creating a document-oriented graph, which, unlike the TF-IDF method, makes it possible to take into account the semantics of sentences and increase the number of stored structural units by 8%. The system of understanding natural language texts for working out and analysis of data is developed.

The method of clustering of texts based on templates is developed. The algorithm of fuzzy duplicate searches in natural texts is constructed. Methods of primary analysis of text and analysis of text blocks are developed

An information and linguistic system for analysis of weakly structured texts in various subject areas were introduced as evidenced by the acts of implementation of the results of the dissertation work. The developed methods for working with semistructured medical data have been tested. Also, the developed methods are used to create a system of work on the summary of hiring workers.

The architecture of the system provides two databases: a document-oriented graph database, which stores data from poorly structured texts, and a database containing general information such as MD5-sums of uploaded files, dictionaries of stop words, keywords, frequency dictionaries, dictionaries for language cliche, metadata, ready-made text templates and sets of pragmatic features (markers). Both databases are used by other modules of the developed system.

Keywords: pragmatic sign, semistructured data, document-oriented graph, NoSQL database, data extraction.

Підписано до друку 06.02.2018.
Формат 60х84/16. Папір офсетний. Друк на різографі.
Ум. друк. арк 1,16. Обл.-вид. арк. 0,9.
Наклад 100 прим. Зам. №28

ТзОВ «Растр-7»
79005, м.Львів, вул. Кн. Романа, 9/1
тел./факс: (032) 235-52-05
Свідоцтво суб'єкта видавничої справи:
ЛВ №22 від 19.11.2002р.