МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили

Шрамов Андрій Віталійович

УДК 004.9

МОДЕЛІ ТА МЕТОДИ АНАЛІЗУ ТЕКСТОВОЇ ІНФОРМАЦІЇ ТА СТВОРЕННЯ СЛОВНИКА ПРЕДМЕТНОЇ ГАЛУЗІ

124 – Системний аналіз

Автореферат
магістерської кваліфікаційної роботи на здобуття освітньої кваліфікації
«Магістр комп’ютерних наук»

Миколаїв – 2021

Магістерська кваліфікаційна робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем.

Науковий керівник: д. т. н., професор, Фісун Микола Тихонович.

Рецензент: к. т. н., доцент, Давіденко Євген Олександрович.

Захист відбудеться 25 лютого 2021 р. о 9^{30} год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.
З дипломною роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «18» лютого 2021 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність роботи на сьогоднішній день словник предметної області – невід'ємний артефакт розробки інформаційних систем. Від того, як оперативно і якісно він буде розроблений багато в чому залежить успішність проекту в цілому. СПО дозволяє досягти взаєморозуміння між замовником і розробником, сформулювати вимоги до програмного продукту (ПП), створити словник даних, проектувати базу даних, створювати призначені для користувача інтерфейси і інструкції, супроводжувати ІС..

Метою кваліфікаційної роботи є аналіз методів побудову словника предметної області та підвищення його повноти у технологіях створення ІС та створення СПО для веб-ресурсу масажного салону шляхом виявлення характеристик термінів, розробки методів та моделей виділення та тлумачення термінів.

Об'єктом дослідження є моделювання предметної області інформаційних систем.

Предметом дослідження є методи, моделі та інформаційні технології автоматизованої побудови словників предметної області для інформаційних систем.

Практичне значення магістерської кваліфікаційної роботи полягає у доведенні отриманих наукових результатів до конкретних технологій, методик, алгоритмів та програмних продуктів. На основі методу виділення багатослівних термінів розроблено алгоритм, який формує терміни та визначає їх частоти в документах.

Апробація результатів дослідження:

Шрамов А. В., Фісун М. Т. «Метод визначення тлумачень (дефініцій) термінів», Інформаційні системи та їх інтелектуалізація: матеріали всеукр. наук.-практ. конф., м. Миколаїв, 9-12 лют. 2021 р. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2021. С. 144-147.

Структура магістерської кваліфікаційної роботи. Магістерська кваліфікаційна робота складається із вступу, 5 розділів, висновків. Загальний обсяг роботи складає 109 сторінок, 54 рисунок, 12 таблиця та 91 посилань на джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** магістерської кваліфікаційної роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У **першому** розділі було розглянуто предметну сферу кваліфікаційної роботи та визначено основні задачі кваліфікаційної роботи.

Словники предметних областей використовуються при створені та проектуванні програмних продуктів під замовлення починаючи з етапу виявлення вимог і закінчуючи супроводом розробки .

На етапі аналізу та збору вимог для розробки деякої інформаційної системи (ІС) словник предметної області дозволяє знайти «спільну мову» між розробником і замовником ІС .

На основі понять (сутностей) визначених у СПО будуються структури баз даних і словники даних .

СПО є необхідним документом для написання інструкцій користувачів і створення користувацьких інтерфейсів .

Оскільки супровід програмних продуктів зазвичай виконується не тими фахівцями, які їх створювали, СПО грає важливу роль для підготовки групи супроводу .

Якщо при створенні СПО будуть зберігатися посилання на джерела виявлення термінів, то СПО може служити основою для створення корпоративних пошукових систем.

На рис. 1.1 показані завдання, при вирішенні яких використовуються словники предметної області.

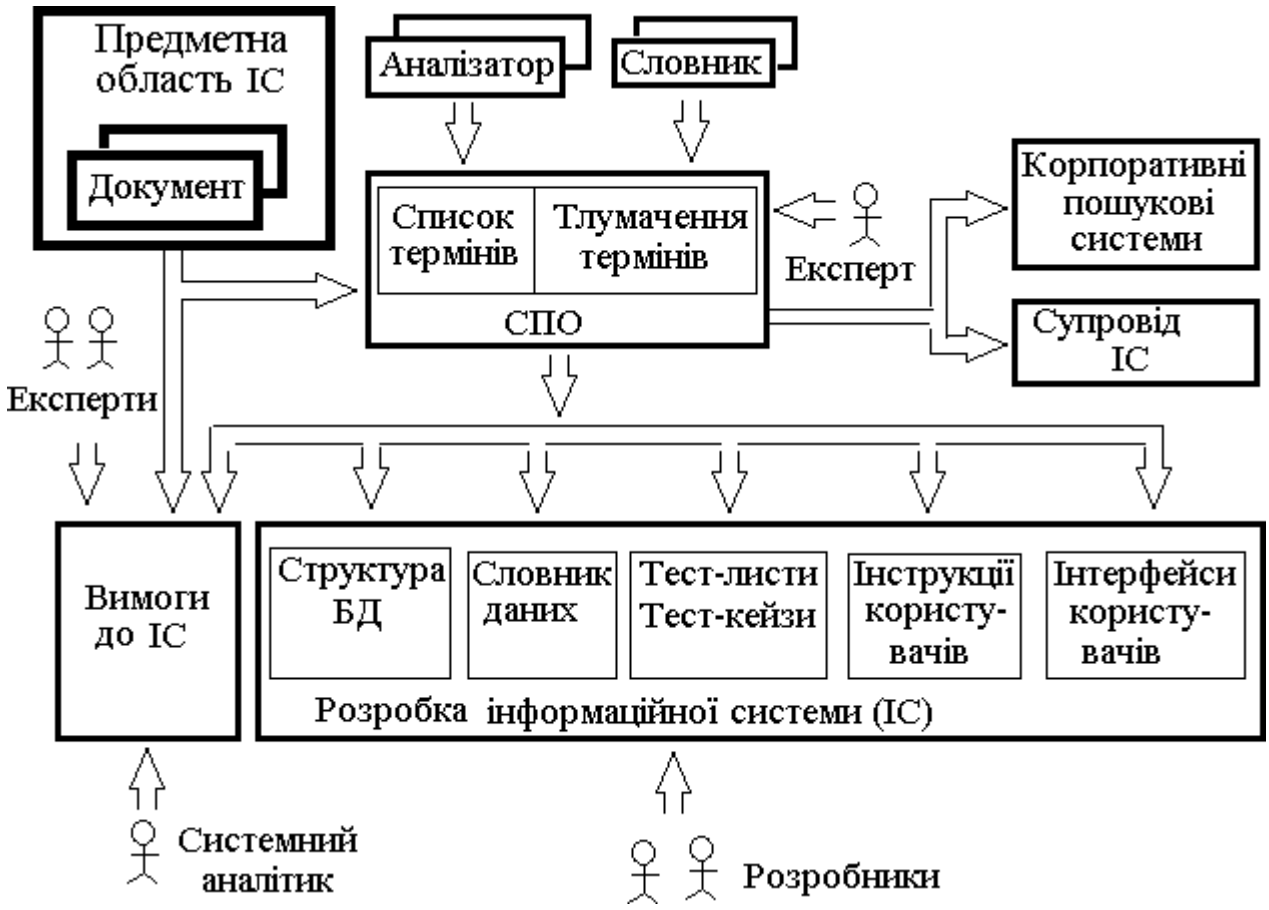

Рис. 1 Загальна схема створення та застосування СПО

Проведено аналіз існуючих засобів морфологічного та синтаксичного аналізу текстів на природній мові.

Основними задачами дипломної роботи є:

– проаналізувати стан моделювання предметної області інформаційних технологій за допомогою СПО;

– проаналізувати метод взаємодії користувача з реляційною базою даних на природній мові;

– удосконалити метод автоматизованого пошуку тлумачення термінів на основі існуючих словників;

– розробити математичну модель представлення терміна, для розробки методів виділення і тлумачення термінів;

– розробити веб-ресурсу масажного салону з використанням

СПО відповідної предметної області.

У **другому** розділі розроблено модель представлення терміну та метод виділення багатослівних термінів та роглянуті основи front-end розробки.

Математична модель терміна Нехай T – деякий текстовий документ, а

$$S = e_1, \ldots, e_l, \ldots, e_m, \quad (1)$$

– деяке речення з документа. Кожни елемент речення e_l може бути окремим словом, або знаком пунктуації. Нехай:

$$tr = e_i, \ldots, e_{i+n}, \quad (2)$$

– термін.

Для представлення терміна необхідно визначити ряд його характеристик:

- елементи, які обов'язково повинні входити в термін;
- можливу кількість елементів в терміні;
- можливі позиції цих елементів в терміні;
- можливі межі терміна.

На думку фахівців – в основі терміна лежить іменник . Назвемо іменник, який несе основне семантичне навантаження на термін – опорним словом. Наприклад, в терміні «реляційна база даних» опорним словом буде іменник «база», а іменник «даних» грає роль визначення для опорного слова. Таким чином, серед елементів, що становлять термін має бути кілька або хоча б один іменників.

Можлива довжина терміна, позиції іменників в терміні, а також їхні межі були визначені експериментальним шляхом. Суть експерименту полягэ у виділенні термінів вручну з текстів, що представляють різні області виробничої діяльності. Враховувалися терміни, що зустрічаються в тесті 2 і більше разів. У кожної предметної області було визначено не менше 100

термінів. Оскільки виявилося, що результати дослідження практично не залежали від предметної області, вони об'єднані в загальну таблицю.

Результати аналізу довжини терміна представлені в таблиці 1.

Таблиця 1.

Імовірність появи терміна певної довжини

Кількість елементів у терміні	Імовірність появи терміна від загальної кількості термінів
1	0.24
2	0.45
3	0.19
4	0.08
5	0.03
6 і більше	0.01

В термін має входити опорне слово – іменник, проте потрібна була інформація про кількість і розташування іменників в терміні. Результати аналізу, наведені в таблицях 2.2 і 2.3 показують, що кількість і позиція іменника в терміні можуть бути будь-якими.

Таблиця 2.

Імовірність появи іменників в терміні

Кількість іменників у терміні	Імовірність появи терміна від загальної кількості термінів	Кількість слів у терміні
1	1	1
2	0.45	2
3	0.55	3
4	0.05	4
5	0.04	5
6 і більше	0.01	6 і більше

Таблиця 3.

Імовірність розташування іменників в терміні

Розташування іменників в терміні	Імовірність появи терміна від загальної кількості термінів
Спочатку	0.49
У середині	0.10
Наприкінці	0.41

Аналіз текстів показав елементи, які задають межі БТ (таблиця 4.).

Таблиця 4.

Можливі межі входження БТ в текст

№	Обмеження зліва	Обмеження справа
1	Пробіл	Пробіл
2	, пробіл	,
3	– пробіл	Пробіл –
4	: пробіл	:
5	; пробіл	;
6	. пробіл	.
7	? пробіл	?
8	! пробіл	!
9	– пробіл	Пробіл –
10	) пробіл	Пробіл (
11	» пробіл	Пробіл «
12	… пробіл	…
13	Займенник пробіл	Пробіл займенник
14	Дієслово пробіл	Пробіл дієслово

Випадок, коли кома входить в термін, виявився єдиним з 1000 проаналізованих термінів – «особи, які приймають рішення». Таким чином, аналіз термінів показав, що вони не містять займенників та дієслів, а також з імовірністю більше ніж 0.99 знаків пунктуації. Це дозволило створити множину елементів, які задають межі терміна.

$$B=\{:,;,.,\&?,!,(,),,,\textit{займеник},\textit{дієслово}\} \quad (3)$$

Оскільки основою для виділення послідовності слів в якості терміна є її повторюваність, необхідно визначати ідентичність термінів в тексті. Для більшості західно-європейських мов, для цього можна використовувати алгоритми нечіткого порівняння рядків. Однак, в слов'янських мовах схиляння термінів по відмінках і зміни, пов'язані з використанням однини чи множини, можуть привести до істотних відмінностей у поданні одного і того

ж терміну. Приклади наведені в таблиці 5. Запропоновано вирішити дану проблему шляхом використання нормалізованої форми подання терміна.

Таблиця 5.

Варіанти представлення терміна

Номер варіанта	Термін		
	Білоруська мова	Українська мова	Російська мова
1	рэляцыйнымі базамі дадзеных	реляційними базами даних	Реляционными базами данных
2	рэляцыйную базу дадзеных	реляційну базу даних	реляционную базу данных
3	рэляцыйная база дадзеных	реляційна база даних	Реляционная база данных
	Нормалізована форма представлення терміна		
	Реляцыйныя база дадзенае	реляційний база дане	реляционный база данное

Таким чином, модель кожного слова *em* має бути представлена словом *e* (безпосередньо з тексту), множиною атрибутів слова *A*, нормалізованою формою представлення *nf*.

$$em = ¿ e, A, nf > ¿ \quad (4)$$

Атрибути необхідні для визначення опорного слова, меж терміну, а також у майбутньому для врахування міжфразових зв'язків.

Метод виділення термінів Метод передбачає чотири етапи. На першому етапі проводиться перетворення форматів документів, виділення однослівних термінів і термінів, які представлені абревіатурою. На другому етапі виконується угруповання малих за обсягом документів для підтримки повноти словника. На третьому етапі виконується виділення багатослівних термінів. На четвертому етапі визначаються міжфразові зв'язки і здійснюється коректування частот появи термінів.

На першому етапі для кожного документу створюється список однослівних термінів (іменників) Ln та список абревіатур La . Кожна запис списку Ln має вигляд em_i= < e, A, nf , q > де q – кількість появ терміну у документі. Якщо слово з тексту $e_j.a_1$ ="noun(іменник)" і знайдено запис у списку$e_i \in \ln \vee e_i.nf = e_j.nf$ то $e_l.q := e_i.q+1$. У протилежному випадку створюється новий запис у Ln.

Список La містить записи виду $r_i = ¿Ab, At>$. де Ab – абревіатура, а At – термін, який вона представляє. З точки зору визначення терміна, відповідного абревіатурі, запропонована наступна класифікація абревіатур.

1. Ініціальна абревіатура мовою документа поміщена в круглі дужки; слова терміна розділені тільки пробілами.

2. Ініціальна абревіатура мовою документа поміщена в круглі дужки; деякі слова терміна об'єднані знаком тире або між словами зустрічається кома, або деякий слова терміна поміщені в круглі дужки.

3. Ініціальна абревіатура з елементами складноскорочених слів на мові документа поміщені в круглі дужки.

4. Ініціальна абревіатура іноземною мовою поміщена в круглі дужки; слова терміна розділені тільки пробілами.

5. Ініціальна абревіатура іноземною мовою поміщена в круглі дужки; деякі слова терміна об'єднані знаком тире або між словами зустрічається кома, або деяке слово терміна поміщено в круглі дужки.

6. Ініціальна абревіатура іноземною мовою поміщена в круглі дужки спільно з багатослівним терміном на іноземній мові.

7. Ініціальна абревіатура іноземною мовою не поміщена в дужки; термін іноземною мовою розташовується перед абревіатурою.

8. Ініціальна абревіатура, поміщена в круглі дужки, іноземною мовою; термін – на мові документа.

9. Неініціальна абревіатура, що представляє хімічну сполуку.

Експериментальним шляхом було визначено ймовірності появи абревіатур в текстах різного змісту таблиця 6. З таблиці видно, що запропоновані типи абревіатур становлять близько 99% від загальної кількості абревіатур.

Таблиця 6.

Імовірність появи абревіатур певного типу у тексті

Тип	1	2	3	4	5	6	7	8	9	Ін.
Імовірність появи	0.44	0.15	0.10	0.13	0.05	0.01	0.03	0.04	0.04	0.01

Нехай S_n представляє деякі речення $S_n = e_1 e_2 \ldots e_i \ldots e_n$ де e_i елемент речення (слово, або розділовий знак). Визначимо слово e як послідовність символів $e = S_1 S_2 \ldots S_j \ldots S_k$ де символ може бути деякою буквою l з множини букв $l \in mL$ або цифрою d із множиною цифр $d \in mD$. Визначимо операцію виділення символу зіслова: $s_j = e[j]$ і відношення приналежності деякого символу до слова: $s_j = e[j]$. Кожна буква характеризується зображенням li, розміром is (рядкова, прописна) і алфавітом al (кирилиця, латиниця), де ls – може приймати два значення low (мала літера) і cap (прописна буква); al – може приймати значення la (латиниця) і ki (кирилиця).

Будемо вважати, що слово має починатися з літери. Визначимо знаки пунктуації, які використовуються всередині речення:

$$P_m = \{:, ;, ., \&, ?, !, (,), ,, -\} \quad (5)$$

Розшифрування абревіатур типів 1 і 2. Вважаючи, що абревіатура може містити тільки букви, запишемо умову першої появи абревіатури в тексті. Нехай $e_m = s1\ s2 \ldots s_j \ldots s_k$ – деяке слово, де m – номер слова, як елемента речення. Якщо, $e_{m-1} = "(" \wedge e_{m+1} = ")" \wedge \forall s_j \mid s_j.ls = cap$ то можна вважати, що e_m є абревіатурою. Позначимо її Ab. Вважаємо, що

текст Та, відповідає Ab розташований зліва від відкриваючої дужки абревіатури і між його елементами відсутні будь-які розділові знаки (тип 1). Тоді:

$$T_a = e_p \dots e_{p+(k-1)} \quad (6)$$

де p = m - (k +1). Операція по визначенню Та буде успішною, якщо:

$$\forall e_i (i = p, p+(k-1)) \vee e_i[1].li = Ab[j].li (j = 1, k) \quad (7)$$

У тесті, який визначає абревіатуру, відповідно до типу В класифікації, може перебувати знак коми. Якщо $e_i \in$ Та ^ e_i =",", то елемент e_i не повинен враховуватисяпри розрахунку p для визначення Та і відповідно до (6) в Та включається слово $T_a = e_p \dots e_{p+(k-1)}$.

Якщо деякий елемент речення e_i містить знак тире (тип 2) – e_i [j] ="-", то його слід вважати двома словами і відповідно зменшити значення p на 1 (p = p -1). Слова, які поміщенні в дужки (перша буква рядкова), пояснюють інші слова терміна не представлені буквами в абревіатурі, тому не повинні враховуватися при розрахунку p для визначення Та. Сформулюємо умову виділення фрагмента тексту, поміщеногов круглі дужки.

Нехай$e_1 e_2 \dots e_i \dots e_j \dots e_n$ являє собою деякий фрагмент тексту, розташований зліва від абревіатури. Тоді якщо e_i ="("^e_j=")"^e_{i+1}[1].is ≠ cap , то всі слова, укладені між e_i і e_j, не повинні враховуватися при розрахунку довжини. Отримуємо p = m - (k +1) + j - i +1.

Розшифрування абревіатур типу 3**.** Якщо умова (7) не дотримується, то можна припустити, що в абревіатурі присутнє складноскорочене слово. Виявлено умови знаходження слів, на основі яких здійснено скорочення, що дозволяє виявити розшифрування абревіатури.

Розшифрування абревіатур типів 4, 5, 6. Терміни для абревіатур типів 4 і 5 виділяються аналогічно виділенню термінів для абревіатур типів 1 і 2.

Для виділення абревіатур типу 6 розроблений алгоритм виділяє текст розшифрування іноземною мовою, його автоматизований переклад та порівняння з імовірним текстом розшифрування на мові документа. При ймовірності збігу текстів 0.8 розшифрування приймається як термін.

Розшифрування абревіатур типів 7 і 8. Для цих абревіатур немає відповідності початкових букв в словах терміна і в абревіатурі. Тому для них прийнято рішення виділяти текст на іноземній мові як термін, що представляє власне ім'я.

Розшифрування абревіатур типу 9**.** Цей тип абревіатури характерний для вузькоспеціалізованих текстів і не вимагає розшифрування.

На другому етапі виконується об’єднання малих за розміром документів з метою підтримки повноти СПО. Виділення багатослівного терміна можливо тільки при його повторному виявленні в тексті. Якщо документ короткий, то повторного входження терміна може не бути. Для визначення впливу розміру документа на якість виділення термінів було проведено дослідження множини документів різного об'єму таблиця 7.

Таблиця 7.

Залежність ймовірності втрати терміна від розміру документа

Розмір документа у словах	Ймовірність втрати терміна
100	0.95
500	0.90
1000	0.55
3000	0.22
5000	0.05
10000	0.03

Нехай $dc_i =< text_i , ds_i >$ деякий документ, де $text_i$ текст документу, ds_i − розмір документу в словах. $Dcs = \{dc_i \mid dc_i .ds \le Ds\ min \wedge dc_i \in Db\}$–множина документів з розміром менше ніж допустимий розмір Ds min . На першому етапі для кожного документу $dc_j \in Dcs$ була визначена множина іменників $Mn_i = \{noun_{i,l}\}$, де $noun =< W, Nq >$ – іменник, e –

слово, Nq – кількість входжень іменника в документ. Для кожного документа $dc_j \in Dcs$ створюється множина унікальних іменників $n1_i = \{noun_{i,k} \mid noun_{i,k} \in Mn_i \wedge Nq_{i,k} = 1\}$, а документ представляється як $dc_i =< text_i , ds_i , Mn_i , Mn1_i >$

Для об'єднання документів введено коефіцієнт близькості K_i ,документа dc_i і dc_j, який визначається двома складовими $(K_{i,j} = K1^1_{i,j} + K1^2_{i,j})$ кількістю іменників, які після об'єднання документів в кластер перестануть бути унікальними $(K1^1_{i,j} = |Mn1_i \cap Mn_j| + |Mn1_j \cap Mn_i|)$ та відносною кількістю співпадаючих іменників:

$$K^2_{i,j} = \frac{|Mn_j|}{|Mn_i \cap Mn_j|} \text{ при } |Mn_i| \geq |Mn_j| , \; K^2_{i,j} = \frac{|Mn_i|}{|Mn_i \cap Mn_j|} \text{ при } |Mn_j| \geq |Mn_i| .$$

Розроблено алгоритм, який на підставі K_i , j дає рекомендації щодо об'єднання документів.

На третьому етапі створюється список багатослівних термінів:

$$Lm = \{r_i\} \; i = 1, n \quad (8)$$

Кожен запис має вигляд:

$$r = ¿ tm, lsn, nft, q > ¿ \quad (9)$$

де tm – множина варіантів представлення терміна, lsn – список опорних слів (іменників), що входять в термін, в нормалізованому вигляді, q – кількість входжень терміна в документ. Однослівні терміни і абревіатури розглядаються як окремі випадки багатослівного терміна.

На четвертому етапі визначаються міжфразові зв'язки для коригування частот появи термінів. Міжфразовий зв'язок (МЗ) означає, що деякий термін в межах одного речення або в наступних реченнях замінюється іншим словом – анафорою. Для визначення впливу МЗ на якість СПО було досліджено 50 науково-популярних, публіцистичних і наукових

текстів обсягом від 2000 до 15000 слів на предмет виявлення типів і ймовірності появи МЗ. В результаті було встановлено, що різні види анафор зустрічаються в текстах різної тематики в середньому 54,3 разів на 1000 слів. При цьому найбільш часто анафора представлена займенником, на другому місці – уточнюючим прикметником, на третьому – порядковим числівником. Зазначені випадки складають близько 90% МЗ.

Представимо аналізований текст dc у вигляді множини речень

$dc = \{S_i\} i = 1, n$, а кожне речення – у вигляді послідовності елементів

(слів і знаків пунктуації). Введемо визначення приналежності терміна реченню $tm_j \in_d S_i$.

Пошук тлумачень суміщений з пошуком термінів. В аналізованому документі можуть вводитися деякі нові поняття (терміни), або приводитися нова інтерпретація відомих. Тоді дефініція терміну може бути включена в текст в безпосередній близькості від самого терміна. У цьому випадку можливо організувати повторний прохід по документу і шукати дефініцію як послідовність слів ліворуч або праворуч від виділеного раніше терміну. Проведений аналіз показав, що існує ряд слів, словосполучень і знаків, які можуть пов'язувати термін з його дефініцією (таблиця. 8).

Таблиця 8.

Словосполучення і знаки, що дозволяють визначити дефініцію

№	Російська мова	Українська мова
1	*tm* «является» *tt*	*to* «є» *tt*
2	*tm* «представляет собой» *tt*	*to* «представляє собою» *tt*
3	*tm* «имеет цель» *tt*	*to* «має на меті» *tt*
4	*tt* «называют» *tm*	*tt* «називають» *to*
5	*tm* «будем понимать» *tt*	*to* «будемо розуміти» *tt*
6	*tm* «который является» *tt*	*to* «який є» *tt*
7	*tm* «обозначает» *tt*	*to* «позначає» *tt*
8	*tm* – *tt*	*to* – *tt*
9	інші	інші

За результатами аналізу частоти використання різних способів зв'язку дефініцій з термінами отримана таблиця 9. Тут номер рядка відповідає номеру рядка в таблиці 8. Таблиця 9 дозволяє зробити висновок про можливість виключити з розгляду випадки 2 і 3 (менше 5% від всіх знайдених дефініцій).

Для перевірки ефективності пошуку тлумачень безпосередньо в тексті, на підставі якого будується СПО, був проведений аналіз документів з предметної області «Електрична сфера в Україні» загальним обсягом 15 тисяч слів. В результаті було виявлено 257 термінів. Для 3% з них відповідно до табл. 8 були знайдені тлумачення. В результаті аналізу документа на 10 тисячі слів з предметної області «Методичні вказівки з інформатики» було виділено 190 термінів. Для 8% з них були знайдені тлумачення безпосередньо в аналізованих текстах. На підставі проведених експериментів зроблено висновок, що пошук дефініцій має певний сенс тільки в текстах, що мають навчальний та рідше науковий характер.

Таблиця 9.

Відсотки появи різних типів зв'язків дефініцій з термінами

Тип зв'язку дефініцій з терміном	Відсоток появи у текстах
1	10
2	4
3	2
4	12
5	10
6	10
7	8
8	43
інші	1

В **третьому** розділі дипломної роботи описано процес розробки проекту, наведено результати роботи.

У **п’ятому** розділі було проаналізовано нормативні документи про охорону праці та визначено умови роботи при написанні дипломної роботи. Описано процес проведення розрахунків для штучного освітлення на робочому місці та зроблено висновки щодо покращення штучного освітлення на робочому місці.

Охорона праці – це система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів, спрямованих на збереження життя, здоров’я і працездатності людини у процесі трудової діяльності.

Законодавство України про охорону праці – це система взаємозв’язаних нормативно-правових актів, що регулюють відносини у галузі охорони праці. Воно складається з Кодексу законів про працю України, Законів України «Про охорону праці», , «Про забезпечення санітарного та епідемічного благополуччя населення», «Про охорону здоров'я», «Про використання ядерної енергії та радіаційний захист», «Про пожежну безпеку», «Про загальнообов’язкове державне соціальне страхування від нещасного випадку на виробництві та професійного захворювання, які спричинили втрату працездатності» та інших. Базується законодавство України про охорону праці на конституційному праві всіх громадян України на належні, безпечні і здорові умови праці, гарантовані статтею 43 Конституції України.

Робота з комп’ютером супроводжується тривалими зоровими навантаженнями і негативно впливає на здоров’я очей. Правильно виконане освітлення робочого місця здійснює позитивний вплив на психофізіологічний стан людини, сприяє підвищенню ефективності та високій працездатності. Досягнення оптимальних умов роботи проводиться шляхом забезпечення природнього освітлення в світлий час доби та сприятливого штучного освітлення в темний період доби.

Для забезпечення умов, що необхідні для зорового комфорту, в системі освітлення повинні бути реалізовані наступні вимоги:

– рівномірне освітлення;

– оптимальна яскравість;

– відсутні відблиски та засвіченості;

– правильний контраст;

– правильна кольорова гамма;

– відсутній стробоскопічний ефект або пульсація світла.

Робота над дипломним проектом відноситься до IV розряду зорових робіт (мінімальний розмір об’єкту, що розглядається – товщина напису букви – 0,3 мм; розряд зорової роботи – робота високої точності) при великому контрасті та світлому фоні (підрозряд зорової роботи «г»).

У **методичній** частині розроблено лабораторні роботи на теми «Основні об’єкти СКБД MS Access. Створення однотабличної БД.» та «Розробка інфологічної моделі та створення структури реляційної БД.».

ЗАГАЛЬНІ ВИСНОВКИ

В роботі вирішена актуальна задача побудови словника предметної області для інформаційних систем на основі розроблених моделей та методів виділення термінів пошуку їх тлумачень. При цьому отримано такі основні результати.

На підставі проведеного аналізу відомих рішень встановлено, що технології побудови словників предметної області мало автоматизовані, немає досліджень щодо виявлення характеристик термінів, Для побудови якісного словника експерт повинен витрачати багато часу.

Розроблено математичну модель представлення терміна, для якої були виконані такі дослідження термінів як ймовірна кількість слів у терміні, кількість та розподіл іменників, визначення обмежень термінів у тексті. Математична модель терміна стала основою для побудови в подальшому методів виділення і тлумачення термінів.

Отримав подальший розвиток метод виділення багатослівних термінів, для якого було розроблено чотири етапи: виявлення однослівних термінів та абревіатур, попереднє групування документів, виявлення багатослівних термінів, коригування частотних характеристик термінів за рахунок виявлення анафор. Це дозволило значно підвищити повноту словника і скоротити час на його побудову.

Проаналізовано метод автоматизованого пошуку тлумачення термінів у словниках завантажених в систему і таких, що знаходяться в Інтернеті. Реалізовано умови вибору тлумачення. Реалізовано алгоритм синтезу тлумачення у випадку коли багатослівний термін не знайдено. Реалізація методу дозволила значно скоротити час роботи експерта зі словником.

Проаналізовано метод спілкування користувача з реляційною базою даних на природній мові, для чого було створено модель реляційної бази даних і відповідний словник, модель запиту користувача, та правила

формування відповіді. Метод дозволив розширити контингент користувачів інформаційної системи, що підвищило продуктивність роботи самої системи та програмістів за рахунок скорочення певного часу.

Розроблено веб-ресурс з використанням СПО відповідної предметної області.

В методичній частині було розроблено лабораторні роботи та наведено повний зміст робіт разом з завданнями для самостійного виконання.

У розділі про охорону праці було досліджено основні нормативні документи, визначено норми умов праці робітника. Проаналізовано умови роботи та порівняно з нормами.

АНОТАЦІЯ

Шрамов А. В.

«Моделі та методи аналізу текстової інформації та створення словника предметної галузі»

Метою кваліфікаційної роботи є аналіз методів побудову словника предметної області та підвищення його повноти у технологіях створення ІС та створення СПО для веб-ресурсу масажного салону шляхом виявлення характеристик термінів, розробки методів та моделей виділення та тлумачення термінів.

Об'єкт дослідження – моделювання предметної області інформаційних систем.

Предмет дослідження – методи, моделі та інформаційні технології автоматизованої побудови словників предметної області для інформаційних систем.

Кваліфікаційна робота містить наступні розділи:

– аналіз предметної сфери. Постановка задачі;
– іформаційні технології для вирішення поставленої задачі;
– програмна реалізація та тестування.

Перший розділ стосується огляду теоретичних відомостей та аналізу існуючих інструментів обробки текстів на природній мові, вимог до програмного забезпечення, що буде реалізовуватись, формулюванню постановки задачі. В *другому* розділі проводиться детальний аналіз методів виділення терміну із текстів на природній мові та тлумачення термінів. В *третьому* розділі наводиться опис розробленого веб ресурсу масажного салону.

Ключові слова: словник предметної області, однослівний термін, багатослівний термін, абревіатура, анафора, міжфразовий зв'язок, словник бази даних, текстовий документ.

ABSTRACT

Shramov A. V.

«Models and methods of text analysis and creating a dictionary of the subject area»

The purpose of the qualification work is to analyze the methods of building a dictionary of the subject area and increase its completeness in the technology of creating IP and creating SPO for the web resource of massage salon by identifying the characteristics of terms, developing methods and models of selection and interpretation of terms.

The object of study - modeling the subject area of information systems.

Subject of research - methods, models and information technologies of automated construction of dictionaries of the subject area for information systems.

Qualification work contains the following sections:

- analysis of the subject area. Formulation of the problem;
- information technology to solve the problem;
- software implementation and testing.

The first section deals with the review of theoretical information and analysis of existing tools for word processing in natural language, the requirements for the software to be implemented, the formulation of the problem. The second section provides a detailed analysis of methods for extracting a term from texts in natural language and interpretation of terms. The third section describes the developed web resource of the massage salon.

Key words: dictionary of subject area, one-word term, multi-word term, abbreviation, anaphora, interphrase connection, database dictionary, text document.