УДК 004.912

# МЕТОД ФОРМУВАННЯ ФРАГМЕНТІВ ТЕКСТУ ДЛЯ ПОШУКОВИХ СИСТЕМ НА ОСНОВІ РОЗПОДІЛУ ТЕРМІНІВ ПО ДОКУМЕНТУ

**О. Б. Кунгурцев, С. В. Ковальчук**
*Одеський національний політехнічний університет*

***Анотація.*** *Розроблено метод, що дозволяє виділити фрагменти тексту на основі аналізу розподілу термінів у відповідь на запит користувача. Описано три підходи встановлення діапазону та формування фрагментів. Представлено алгоритм формування фрагментів. Описано принцип вибору діапазону для кожного терміну. Представлено алгоритм формування фрагментів для конкретного терміну. Також описано принцип порівняння декількох фрагментів між собою.*

***Ключові слова:*** *діапазон, фрагмент, термін, природна мова, інформаційна система, предметна область.*

## Постановка проблеми

Збільшення обсягу інформації, необхідної людям для їх діяльності в самих різних областях, зростає дуже швидко. Традиційних методів пошуку документів з використанням відомих класифікацій, наприклад УДК, явно недостатньо. Безперервно удосконалюються способи пошуку за ключовими словами [1], але і вони мають ряд недоліків (суб'єктивізм у визначенні ключових слів, відсутність їх відносної «значущості» для документа). Зміст документа можна уявити більш точно, якщо замість ключових слів використовувати терміни, які визначаються під час попередньої обробки тексту в автоматизованому режимі [2]. Частота появи певних термінів може бути однією з характеристик, що визначають зміст документа.

Знаходження актуальної для користувача інформації є головною метою будь якої інформаційної системи (ІС) [3]. Запити до ІС на природній мові (ПМ) можуть бути різної структури і змісту. Запит може містити в собі інформацію про одне ключове поняття або декілька, які можуть бути взаємно пов'язані між собою [4]. Документи можуть бути досить великих розмірів і користувачеві на пошук потрібного фрагмента тексту, в рамках такого документа, також може знадобитися багато часу. Відомі рішення не дозволяють виділити найбільш «відповідний» фрагмент тексту.

Таким чином формується загальна проблема, яка полягає в великій траті часу на пошук потрібної інформації в документо-орієнтованих ІС через відсутність механізму виділення окремих текстів та їхніх фрагментів, що враховує розподіл термінів по тексту, а також взаємний розподіл термінів, надає користувачеві можливість отримання «пробних» фрагментів документа.



## 1. Аналіз останніх досліджень та публікацій

Питання щодо створення інтерфейсу для спілкування користувачів з ІС на ПМ неодноразово піднімалися і обговорювались [5]. Але не було детально описано в якому вигляді і яким чином ІС буде представляти результат виконання запиту користувача. Також існують праці, на основі яких будувався словник предметної області з урахування позицій термінів в документі [2], але не розглядалося питання щодо виявлення фрагментів тексту з вказаними позиціями їхнього розташування.

Не завжди потрібно просто повертати увесь документ у відповідь на запит користувача. Іноді потрібно давати фрагменти тексту, які максимально задовольняють поставленому запиту до ІС. В кінцевому результаті, користувач сам приймає рішення, чи задовольняє його результат і чи даний документ йому потрібний.

Формування фрагментів документу, які відповідають критерію пошуку, є основним завданням для коректної відповіді ІС на запит користувача.

На підставі аналізу загальних тверджень [6], текст характеризувався тільки кількістю входжень ключових слів в тексті. Не можна було виділити фрагмент тексту найбільш інтенсивно використовуваного терміну. Також не можна було виділити окремий документ, або фрагмент документу, в якого існує певне поєднання термінів.

Терміни в документах розташовані неоднорідно. В деяких місцях документа термін може зустрічатися частіше, а в деяких і взагалі не зу-

стрічатися. Таким чином кожен термін має своє розподілення по документу [7].

Це все приводить до того, що потрібно не тільки знати кількість термінів в тексті, але і їхній розподіл по тексту. Наступна проблема - формування фрагментів у зв'язку з прийняттям рішення, що вносити в фрагмент а що ні. Сучасні рішення пропонують ряд методів пошуку текстової інформації для знаходження позицій ключових слів [8], та дають рекомендації по розподіленню ключових слів по сторінках інтернет сайтів [9]. Значну кількість робіт було присвячено логіко-лінгвістичній моделі, а саме порівнянню текстів за змістом [10, 11]. Дані рішення не вирішують проблему формування та швидкого пошуку фрагментів тексту для конкретного терміну. Таким чином, існує проблема виділення фрагментів тексту певної тематики, заданої користувачем у вигляді одного або декількох термінів.

## 2. Постановка задачі

Для формування фрагментів документу для кожного терміну з урахуванням їхньої позиції в тексті та можливості порівняння фрагментів різних термінів між собою потрібно вирішити наступні завдання.

1. Проаналізувати вхідний текст з метою виявлення розподілу кожного терміну.
2. Визначити відносну інтенсивність появи терміну в документі.
3. Визначити терміни, які рівномірно розподілені по документу і терміни, які мають нерівномірний розподілу по документу.
4. Сформувати фрагменти для кожного терміну.
5. Знайти фрагменти в яких сумісно зустрічаються декілька термінів.

В результаті вирішення виділених завдань буде сформовано фрагменти тексту для кожного терміну, який буде містити в собі інформацію про розташування фрагменту в тексті та кількості входжень конкретного терміну в даний діапазон. Також буде надана можливість порівняння між собою різних фрагментів кожного терміну, для досягнення високої точності в формуванні результату на запит користувача.

## 3. Мета статті

Головною метою цієї роботи є розробка методу формування фрагментів документу для кожного терміну та можливість порівнювати різні фрагменти між собою, що дасть змогу зменшити час на виконання запиту користувача, збільшити точність пошуку та надати користувачеві найбільш підходящі фрагменти.

## 4. Висвітлення основних понять

Фрагменти тексту певного терміну формуються на основі розподілення терміну по документу. Таким чином можна виділити фрагменти з найбільшою інтенсивністю терміну і формувати відповідь на запит користувача у вигляді рейтингу перших фрагментів, які найбільш задовольняють запит до ІС. Для одного терміну не виникає проблем формування відповіді за запит користувача, але при необхідності виявлення спільної частини між фрагментами різних термінів з'являється неоднозначність, пов'язана з вибором спільної межі, яка буде задовольняти умовам запиту користувача.

Пошук спільних меж пересікання фрагментів різних термінів потрібно здійснювати в реальному часі, так як варіацій порівнянь фрагментів між собою дуже багато і обробити їх усіх занадто затратно по часу, і для зберігання такої інформації потрібні великі затрати пам'яті ІС.

Дане дослідження основується на праці [2] і є наступним етапом в даній предметній області. Основні поняття, які розглядаються в даному дослідженні це фрагменти тексту та діапазон речень. Діапазон речень – це кількісна оцінка, яка містить в собі певну кількість речень. Кожен термін має певну кількість повторень і відносно цієї кількості повторень і буде формуватися діапазон. Розглядаючи якийсь конкретний термін, можна сказати, що він розташований по тексту не рівномірно і зазвичай зустрічається не більше одного разу в реченні. Тому діапазони служать для того, щоб захопити декілька повторень конкретного терміну. Чим більший діапазон, тим більша ймовірність того, що в ньому більша кількість повторень обраного терміну. Для кожного терміну сформований діапазон має однаковий розмір і виступає в ролі шкали. Поняття «фрагмент» представляє собою певну кількість діапазонів. Для кожного терміну можуть бути фрагменти різного розміру. Фрагменти формуються на основі аналізу діапазонів.

В даній роботі будуть розглядатися 3 випадки з вибором розміру діапазону, і результатів даного вибору. Всі підходи мають певні спільні принципи, а також індивідуальні відмінності. Кожен з трьох випадків має як певні переваги так і недоліки над іншими описаними підходами.

Кожен термін має свою інтенсивність появи в тексті. Для кожного документу може бути значна кількість термінів. Відповідно терміни з великою кількість повторень мають більшу вагомість для тексту. Якщо потрібний термін зустрічається, в аналізованому документі, досить рід-

ко, то потрібно пропустити даний документ і проаналізувати його на більш пізніх стадіях пошуку.

Якщо термін розташований рівномірно по тексту, то це означає, що не можна виділити підходящі фрагменти, так як увесь текст може бути в рамках одного фрагменту. В такому випадку результатом буде увесь документ. У випадку сумісних фрагментів між різними термінами, термін з рівномірним розподілом не буде впливати на формування фрагментів. Рівень рівномірності розподілу термінів по документу $G$ можна представити наступним чином:

$$G = \frac{\sum_{q=1}^{size}(H_{sr} - H_q)^2}{size}, \quad (1)$$

де $H_{sr}$ – середня кількість повторень терміну в кожному діапазоні,

$H_q$ – фактична кількість повторень терміну в діапазоні,

$size$ – кількість діапазонів.

Середня кількість повторень терміну $H_{sr}$ має наступний вигляд:

$$H_{sr} = \frac{\sum_{q=1}^{size} H_q}{size}. \quad (2)$$

Якщо в результаті аналізу термін виявився з нерівномірним розподілом, то можна застосувати для подальшого аналізу один з представлених нижче підходів.

### *4.1. Для кожного терміну встановлюється унікальний діапазон речень*

У даному випадку кожен термін $tx$ конкретного тексту $T$ характеризується своїм діапазоном речень**.** Встановлені діапазони розраховуються і є актуальними тільки для конкретного тексту $T$.

Середня довжина речення $ASL$ характеризується наступним чином:

$$ASL = \frac{allWords}{SCount}, \quad (3)$$

де $allWords$ – загальна кількість слів в тексті $T$,

$SCount$ – кількість речень в тексті $T$.

Розподілення термінів по тексту різноманітне, і як правило, термін зустрічається в реченні не більше 1-2 разів. Тому потрібно вирішити, яку кількість термінів очікувати на заданий діапазон.

Формула, яка характеризує текст $T$ відносно конкретного терміну $tx$, має наступний вигляд:

$$Y = \frac{txCount}{SCount} * \Delta, \quad (4)$$

де $txCount$ – кількість повторень терміну $tx$ в тесті $T$,

$\Delta$ – діапазон речень для конкретного терміну $tx$,

$Y$ – очікувана кількість терміну $tx$ в кожному діапазоні.

Очікувану кількість терміну в кожному діапазоні $Y$ можна встановити будь-яку у відповідності з поставленою задачею і цілями. Але також дану величину можна спробувати представити у вигляді загальної рекомендації, яка говорить скільки можна очікувати термінів на відрізку даного діапазону в текстах технічного характеру, якщо діапазон становить одне речення.

Очікувану кількість терміну в кожному діапазоні $Y$ можна представити в наступним чином:

$$Y = \frac{ASL}{ASLflash} * Ykof, \quad (5)$$

де $ASLflash$ – середня довжина речень в тексті $T$ по методології індексу Флеша,

$Ykof$ – коефіцієнт очікуваної кількості термінів.

Індекс Флеша $FRE$ для української мови представляється наступним чином [12]:

$$FRE = 206.835 - (1.3 * ASL) - (60.1 * ASW), \quad (6)$$

де $ASW$ – середня кількість складів в слові.

Середню кількість складів в слові можна представити наступним чином:

$$ASW = \frac{allSullables}{allWords}, \quad (7)$$

де $allSullables$ – загальна кількість складів в тексті $T$.

Індекс Флеша (6) розподіляється наступним чином:

$FRE = 100$ – легкий для сприйняття текст, середня довжина речень не більше 12 слів;

$FRE = 65$ – середньої важкості для сприйняття текст, середня довжина речень від 15 до 20 слів;

$FRE = 30$ – важкий для сприйняття текст, середня довжина речень до 25 слів;

$FRE = 0$ – дуже важкий для сприйняття текст, середня довжина речень до 37 слів.

Щоб визначити діапазон речень $\Delta$ для терміну $tx$, потрібно прирівняти формули (4) та (5). Після математичних обрахунків, отримаємо діапазон речень $\Delta$ в наступному вигляді:

$$\Delta = \frac{allWords * Ykof}{ASLflash * txCount}. \quad (8)$$

В даному випадку, з урахування предметної області (наукові статті та книги), встановлюємо середню довжину речень по Флешу наступним чином: $ASLflash = 20$, це означає, що тексти рахуються середньої важкості для сприйняття ($FRE = 65$). Коефіцієнт очікуваної кількості термінів $Ykof = 1$, так як в даному випадку очікується більш детальні діапазони і відповідно більша їхня кількість. Таким чином формули (4, 5) можна представити наступним чином:

$$\Delta = \frac{allWords}{20 * txCount}. \quad (9)$$

Сформувавши діапазони $\Delta$, для кожного терміну $tx$, потрібно занести кількість повторень кожного терміну для свого діапазону.

Спочатку потрібно знайти в яких реченнях зустрічається потрібний термін $tx$. Для цього потрібно звернутися до попередніх робіт [2] в яких описано процес знаходження термінів і їхніх позицій в тексті. На основі даних результатів можемо представити для кожного терміну $tx$ таблицю, яка містить в собі інформацію про номер речення та кількість повторень конкретного терміну в даному реченні: $txTable = \{txFind_j\} j = 1, txTableCount$. На початку $txTable = \varnothing$. Кожен запис представимо у вигляді кортежу:

$$txFind_j \in tx_i =< Snum_j, findCount_j >, \quad (10)$$

де $Snum_j$ – номер речення, в якому зустрівся термін $tx_i$,

$findCount_j$ – кількість повторень терміну в конкретному реченні.

Провівши ряд експериментів, можна сказати, що в переважній кількості текстів, терміни зустрічають в реченні не більше 1-2 разів, як правило один раз в реченні. Для прикладу (Технічний текст), приведений невеликий технічний текст «Комп'ютерні системи» $T$ з наступними характеристиками: $allWords = 5553$, $SCount = 295$, $tx_i = система$, $txCount_i = 64$. Даний термін $tx_i$ має найбільшу кількість появ в тексті $T$. Представимо результати наступними парами $< Snum_j, findCount_j >$: <4,**2**>, <5,1>, <12,1>, <13,1>, <14,1>, <16,1>, <17,1>, <19,1>, <21,1>, <26,1>, <48,1>, <52,1>, <53,1>, <54,1>, <56,1>, <57,1>, <59,1>, <60,1>, <65,1>, <66,1>, <70,1>, <86,1>, <88,1>, <89,1>, <95,1>, <96,1>, <98,1>, <101,1>, <102,1>, <109,1>, <110,1>, <111,1>, <112,1>, <115,1>, <120,1>, <121,1>, <122,1>, <125,1>, <129,1>, <130,1>, <133,1>, <134,**2**>, <136,1>, <139,1>, <146,1>, <147,1>, <156,1>, <157,1>, <161,**3**>, <175,1>, <176,1>, <200,1>, <216,1>, <222,1>, <224,1>, <235,1>, <273,1>, <290,1>, <291,1>.

Наступним кроком потрібно занести термін $tx_i$ до вказаного діапазону, тобто підрахувати кількість повторень терміну для кожного діапазону. Для того, щоб розглядати діапазон не просто як кількість речень, потрібно розширити його визначення як розширений діапазон, і це поняття буде характеризувати вже конкретну ділянку тексту. Множину розширених діапазонів представимо наступним чином: $\Delta Table = \{\Delta Value_d\} d = 1, \Delta TableCount$. Кожен розширений діапазон представимо у вигляді кортежу:

$$\Delta Value_d =< \Delta Len_d, \Delta Count_d >, \quad (11)$$

де $\Delta Len_d$ – розташування діапазону (містить в собі початок та кінець діапазону),

$\Delta Count_d$ – кількість входжень терміну $tx_i$ в діапазон.

Представимо місце знаходження діапазону $\Delta Len_d$ у вигляді кортежу:

$$\Delta Len_d =< \Delta S\min_d, \Delta S\max_d >, \quad (12)$$

де $\Delta S\min_d$ – номер речення початку діапазону $\Delta Value_d$,

$\Delta S\max_d$ – номер речення кінця діапазону $\Delta Value_d$.

Кінець діапазону рахуємо наступним чином:

$$\Delta S\max_d = d * \Delta. \quad (13)$$

Початок діапазону рахуємо наступним чином:

$$\Delta S\min_d = \Delta S\max_d - \Delta + 1. \quad (14)$$

Перейдемо безпосередньо до внесення кількості терміну $tx_i$ до кожного розширеного діапазону $\Delta Value_d$.

Якщо виконується умова:

$$(Snum_j >= \Delta S\min_d) \wedge (Snum_j <= \Delta S\max_d), \quad (15)$$

то виконуємо наступні дії: $\Delta Count_d = \Delta Count_d + findCount_j$.

Якщо $Snum_j > \Delta Snum_d$, то переходимо до наступного діапазону $\Delta Value_{d+1}$.

Після виконаних дій ми отримуємо кількість повторень терміну $tx_i$ для усіх діапазонів $\Delta Value_d$.

Розглянемо (Технічний текст) в якості прикладу. Для терміну $tx_i = система$ діапазон становить $\Delta = 6$. Представимо діапазони наступними парами $< \Delta Len_d, \Delta Count_d >$: <(1-6),3>, <(7-12),1>, <(13-18),4>, <(19-24),2>, <(25-30),1>, <(31-36),0>, <(37-42),0>, <(43-48),1>, <(49-54),3>, <(55-60),4>, <(61-66),2>, <(67-72),1>, <(73-78),0>, <(79-84),0>, <(91-96),2>, <(97-102),3>, <(103-108),0>, <(109-114),4>, <(115-120),2>, <(121-126),4>, <(127-132),2>, <(133-138),4>, <(139-144),1>, <(145-150),2>, <(151-156),1>, <(157-162),4>, <(163-168),0>, <(169-174),0>, <(175-180),2>, <(181-186),0>, <(187-192),0>, <(193-198),0>, <(199-204),1>, <(205-210),0>, <(211-216),1>, <(217-222),1>, <(229-234),0>, <(235-240),1>, <(241-246),0>, <(247-252),0>, <(253-258),0>, <(259-264),0>, <(265-270),0>, <(271-276),1>, <(277-282),0>, <(283-288),0>, <(289-294),2>, <(295-300),0>.

Наступним кроком потрібно сформувати фрагменти тексту, які будуть містити в собі певну кількість терміну $tx_i$. Математичне представлення формування фрагментів (де $m$ - кількість діапазонів) має наступний вигляд:

$$F = \{((\Delta Value_d \mid \Delta Len_d) \geq m) \wedge \exists(((\Delta Value_f \mid \Delta Len_d) = \Delta Len_{f+1}) \vee \Delta Len_d = \Delta Len_{f-1})\}d,\ f = 1, m. \quad (16)$$

Множину фрагментів можна представити наступний чином: $FragTable = \{txFrag_z\} z = 1, FragTableCount$. Кожен фрагмент буде представляти собою кортеж:

$$txFrag_z = < txFragLen_z, countFrag_z >, \quad (17)$$

де $txFragLen_z$ – розташування фрагменту (містить в собі початок та кінець фрагменту),

$countFrag_z$ – кількість входжень терміну $tx_i$ в фрагмент.

Представимо місце знаходження фрагменту $txFragLen_z$ у вигляді кортежу:

$$txFragLen_z = < txFragS\min_z, txFragS\max_z >, \quad (18)$$

де $txFragS\min_z$ – номер речення початку фрагмента $txFrag_z$,

$txFragS\max_z$ – номер речення кінця фрагмента $txFrag_z$.

Знаходження фрагментів представлено у вигляді схеми алгоритму на рис. 1.

Розглянемо певний текст (Науково-технічний текст) в якості прикладу. Представимо фрагменти наступними парами $< txFragLen_z, countFrag_z >$: <(1-30),11>, <(43-72),10>, <(85-162),32>, <(175-180),2>, <(199-204),1>, <(211-228),3>, <(235-240),1>, <(271-276),1>, <(289-294),2>.

Порівнювати фрагменти між різними термінами можна по принципу пересічення множин, тобто результатом пересікання фрагменту $F1$ з фрагментом $F2$ двох термінів буде новий спільний фрагмент $Fres$. Загальна формула має наступний вигляд:

$$Fres = F1 \cap F2 = \{c \mid c \in F1 \wedge c \in F2\} \quad (19)$$

Порівнювати фрагменти між різними термінами потрібно кожен з кожним, так як їхні розміри і розташування можуть бути різні. Спільні фрагменти між різними термінами будуть представляти собою масив записів: $FTCommon = \{txFragCommon_u\} u = 1, FTCount$. Кожен спільний фрагмент буде представляти собою кортеж:

$$txFragCommon_u = < txFragCLen_u, countFragC_u >, \quad (20)$$

де $txFragCLen_u$ – розташування фрагменту (містить в собі початок та кінець фрагменту),

$countFragC_u$ – список з термінами і їхньою кількістю повторень у фрагментах, які порівнюються. Представимо місце знаходження спільного фрагменту $txFragCLen_u$ у вигляді кортежу:

$$txFragCLen_u = < txFragCS\min_u, txFragCS\max_u >, \quad (21)$$

де $txFragCS\min_u$ – номер речення початку фрагмента $txFragCommon_u$,

$txFragCS\max_u$ – номер речення кінця фрагмента $txFragCommon_u$.

Список з термінами має наступний вигляд: $countFragC_u = \{coutFtx_t\} t = 1, sizeFC$,

$sizeFC$ – кількість термінів в спільному фрагменті. Кожен запис $coutFtx_t$ представляє наступний кортеж:

$$coutFtx_t = < tx_i, cuurentCountFtx_i >, \quad (22)$$

де $cuurentCountFtx_i$ – кількість повторень терміну $tx_i$ в спільному фрагменті.

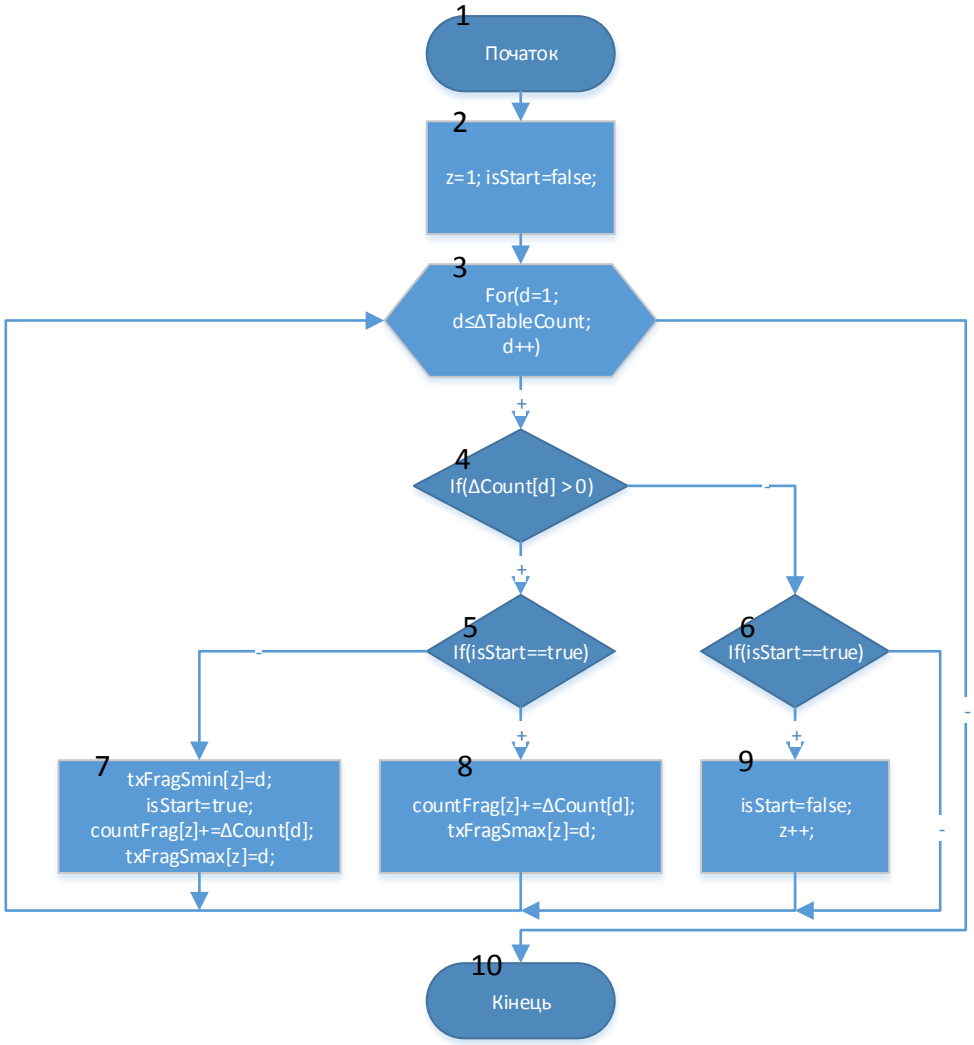


Рис. 1. Схема алгоритму формування фрагментів тексту

Таким чином при порівнянні декількох фрагментів різних термінів $tx_i$, і знайшовши їхню спільну зону, маємо інформацію про початок і кінець спільного фрагменту, та інформацію про терміни і їхню кількість повторень в спільному фрагменті.

Між собою порівнюються одночасно тільки два фрагменти різних термінів. Результатом такого порівняння є новий фрагмент який можна порівнювати з фрагментом третього терміну і т.д.

Представимо, що позиції фрагменту першого терміну наступні: $txFragS\min_z \in tx_i$, $txFragS\max_z \in tx_i$. Тоді позиції фрагменту другого терміну наступні: $txFragS\min_p \in tx_{i+1}$, $txFragS\max_p \in tx_{i+1}$.

Для того, щоб знайти спільну зону двох фрагментів, потрібно виконати одну з наступних умов:

Якщо виконується умова $b1$:

$$b1 = (txFragS\min_z \le txFragS\min_p) \wedge (txFragS\max_z \ge txFragS\max_p), \quad (23)$$

то позиції спільного фрагменту наступні: $txFragCS\min_u = txFragS\min_z$, $txFragCS\max_u = txFragS\max_z$.

Якщо виконується умова $b2$:

$$b2 = (txFragS\min_z \geq txFragS\min_p) \wedge (txFragS\max_z \leq txFragS\max_p), \quad (24)$$

то позиції спільного фрагменту наступні: $txFragCS\min_u = txFragS\min_p$, $txFragCS\max_u = txFragS\max_p$.

Якщо виконується умова $b3$:

$$b3 = (txFragS\min_z \leq txFragS\min_p) \wedge (txFragS\max_z \leq txFragS\max_p), \quad (25)$$

то позиції спільного фрагменту наступні: $txFragCS\min_u = txFragS\min_z$, $txFragCS\max_u = txFragS\max_p$.

Якщо виконується умова $b4$:

$$b4 = (txFragS\min_z \geq txFragS\min_p) \wedge (txFragS\max_z \geq txFragS\max_p), \quad (26)$$

то позиції спільного фрагменту наступні: $txFragCS\min_u = txFragS\min_p$, $txFragCS\max_u = txFragS\max_z$.

В даному випадку виникає проблема підрахунку кількості термінів в спільному фрагменті, так як кожен термін має різні діапазони. Тому потрібно перерахувати діапазон для терміну, який наразі має більшу кількість речень в діапазоні. Перерахувати потрібно так, щоб діапазони були однакового розміру, для цього використовуємо формули (9 – 15).

При умові, що діапазони однакові, можна здійснювати підрахунок кожного термінів в спільному фрагменті. Для кожного терміну $tx_i$, який порівнюється здійснюється підрахунок повторень відносно діапазонів, які увійшли у спільний фрагмент.

### *4.2. Для усіх текстів і термінів діапазон однаковий*

Потрібно визначити $\Delta$ який буде характеризувати усі тексти.

Спочатку вияснимо якої величини тексти буде характеризувати даний діапазон $\Delta$. Максимальне значення можна опустити, так як тексти можуть бути достатньо великих розмірів. Отже, до уваги беруться тексти малого розміру і потрібно встановити мінімальний розмір тексту для виявлення діапазону $\Delta$.

Мінімальним розміром будемо приймати наукові тези об'єм яких може сягати від 500 слів. *Індекс Флеша* в текстах даного типу в середньому складає **65** і це становить **20** слів на одне речення. Отже, розмір мінімального тексту буде складати **25** речень.

Наступним кроком потрібно провести дослід, на основі якого можна сформувати $\Delta$.

Потрібно проаналізувати наукові тези розміром в середньому 25 речень (500 слів) і виявити в них терміни, які найчастіше зустрічаються і відповідають суті теми. Кожен текст аналізується синтаксичним аналізатором [13], який був описаний в роботі [2]. На виході отримуємо список усіх термінів з кількістю повторень кожного. В якості експерта, обираємо 3-4 терміни які найчастіше зустрічаються і відповідають темі. Після цього знаходимо середнє значення повторень. Формуємо для кожного тексту такі дані як $allWords$ – кількість слів в тексті та $txCount$ – середня кількість повторень терміну. Відповідно до формули (9) рахуємо $\Delta$ для кожного тексту (обрахунки заокруглюємо в більшу сторону). Результати аналізу 50 тез можна представити наступними парами $< allWords, txCount, \Delta >$: <567,16,2>, <500,9,3>, <509,5,5>, <599,13,3>, <553,12,3>, <645,8,4>, <555,7,4>, <416,7,3>, <538,12,3>, <487,8,3>, <537,10,3>, <470,5,5>, <554,10,3>, <430,7,3>, <610,9,4>, <570,9,4>, <501,11,3>, <490,15,2>, <530,8,4>, <515,9,3>, <470,5,5>, <545,8,4>, <517,10,3>, <601,12,3>, <598,13,3>, <537,7,4>, <505,7,4>, <477,5,5>, <508,8,4>, <543,9,3>, <613,13,3>, <546,12,2>, <518,9,3>, <616,9,4>, <523,10,3>, <583,11,3>, <577,7,5>, <493,7,4>, <471,8,3>, <555,10,3>, <519,15,2>, <435,5,5>, <575,8,4>, <619,10,4>, <573,8,4>, <527,11,3>, <495,7,4>, <570,13,3>, <640,11,3>, <530,8,4>.

Таким чином на основі даного експерименту можна сказати, що середнє значення діапазону становить $\Delta = 4$.

Занесення терміну $tx_i$ до вказаного діапазону відбувається за допомого описаного вище методу та формул: (9 – 15).

Розглянемо попередній текст (Науково-технічний текст) в якості прикладу. Для терміну $tx_i = система$ новий діапазон становить $\Delta = 4$. Представимо діапазони наступними парами $< \Delta Len_d, \Delta Count_d >$: <(1-4),2>, <(5-8),1>, <(9-12),1>, <(13-16),3>, <(17-20),2>, <(21-

24),1>, <(25-28),1>, <(29-32),0>, <(33-36),0>, <(37-40),0>, <(41-44),0>, <(45-48),1>, <(59-52),1>, <(53-56),3>, <(57-60),3>, <(61-64),1>, <(65-68),2>, <(69-72),1>, <(73-76),0>, <(77-80),0>, <(81-84),0>, <(85-88),2>, <(89-92),1>, <(93-96),2>, <(97-100),1>, <(101-104),2>, <(105-108),0>, <(109-112),4>, <(113-116),1>, <(117-120),1>, <(121-124),2>, <(125-128),1>, <(129-132),2>, <(133-136),4>, <(137-140),1>, <(141-144),0>, <(145-148),2>, <(149-152),0>, <(153-156),1>, <(157-160),1>, <(161-164),3>, <(165-168),0>, <(169-172),0>, <(173-176),2>, <(177-180),0>, <(181-184),0>, <(185-188),0>, <(189-192),0>, <(193-196),0>, <(197-200),1>, <(201-204),0>, <(205-208),0>, <(209-212),0>, <(213-216),1>, <(217-220),0>, <(221-224),2>, <(225-228),0>, <(229-232),0>, <(233-236),1>, <(237-240),0>, <(241-244),0>, <(245-248),0>, <(249-252),0>, <(253-256),0>, <(257-260),0>, <(261-264),0>, <(265-268),0>, <(269-272),0>, <(273-276),1>, <(277-280),0>, <(281-284),0>, <(285-288),0>, <(289-292),2>, <(293-296),0>.

Фрагменти тексту формуємо на основі формул (17), (18) та алгоритму зображеному на рис. 1.

Розглянемо (Технічний текст) в якості прикладу. Представимо фрагменти наступними парами $< txFragLen_z, countFrag_z >$: <(1-28),11>, <(45-72),12>, <(85-104),8>, <(109-140),15>, <(145-148),2>, <(153-164),5>, <(173-176),2>, <(197-200),1>, <(213-216),1>, <(221-224),2>, <(233-236),1>, <(273-276),1>, <(289-292),2>.

Порівнювати фрагменти між різними термінами можна по принципу пересічення множин як описано вище, за допомогою формул: (19– 26).

В даному випадку не виникає проблема підрахунку кількості термінів в спільному фрагменті, так як усі діапазони однакові.

### *4.3. Для усіх текстів і термінів діапазон однаковий і становить одне речення*

В даному підході вже відомо що розмір діапазону становить одне речення $\Delta = 1$. Таким чином вносити терміни $tx_i$ в діапазони не потрібно, так як це виконується на ранніх етапах описаних в першому підході за допомогою формули (10).

В даному підході потрібно ввести таке поняття як *відстань між реченнями певного логічного змісту* $betweenSCount$. Міжфразові зв'язки були описані в роботі [14]. Також в праці [15] було описано спосіб виявлення логічних зв'язків між частинами текстових документів. Це означає, що між реченням, які несуть певну інформацію про термін, може бути розрив у вигляді речень, які не мають даний термін. Від величини даного розриву залежить чи відносяться речення без терміну, між реченнями, які мають даний термін, до логічного змісту даного фрагменту тексту.

На основі загальних тверджень будемо вважати, що розрив розміром в три речення є допустимим і по змісту відноситься до речення перед ним з терміном і реченням після нього з терміном. Таким чином $betweenSCount = 3$.

Фрагменти тексту формуємо на основі формул (17), (18) та алгоритму зображеному на рис. 1. При використанні алгоритму потрібно дещо змінити умову в блоці під номером 6. Для даного методу умова в даному блоці буде мати наступний вигляд: *if(isStart==true && d>(txFragSmin[z]+betweenSCount))*. Це дасть змогу зносити в фрагмент речення, які входять в допустимий інтервал між реченнями певного логічного змісту.

Розглянемо попередній текст (Науково-технічний текст) в якості прикладу. Представимо фрагменти наступними парами $< txFragLen_z, countFrag_z >$: <(4-5),3>, <(12-21),7>, <(26-26),1>, <(48-60),8>, <(1-28),11>, <(65-70),3>, <(86-89),3>, <(95-102),5>, <(109-115),5>, <(120-139),11>, <(146-147),2>, <(156-161),5>, <(175-200),3>, <(216-216),1>, <(222-224),2>, <(235-235),1>, <(273-273),1>, <(290-291),2>.

Порівнювати фрагменти між різними термінами можна по принципу пересічення множин як описано вище, за допомогою формул: (19 – 26).

В даному випадку не виникає проблема підрахунку кількості термінів в спільному фрагменті, так як усі діапазони однакові.

### *4.4. Експеримент*

Виявлення фрагментів є більш ефективним способом ніж надання усього тексту. Чим менше тексту виділено тим ефективнішим є результат при умові, що виділені фрагменти відповідають змісту документа. Залежність розмірів фрагментів певного терміну до об'єму всього документу можна представити наступним чином:

$$Ef_i = \frac{SCount - \sum_{z=1}^{n} txFragLen_z}{SCount} * 100\%, \quad (27)$$

де $Ef_i$ – ефективність використання фрагментів виражена у відсотках для *i*-того терміну,

$SCount$ – кількість речень в документі,

$\sum_{z=1}^{n} txFragLen_z$ – число речень, які потрапили у фрагменти *i*-того терміну.

Ефективність для декількох термінів, в рамках одного фрагменту, обраховується спочатку для кожного терміну окремо, після чого знаходиться середнє арифметичне серед усіх термінів в фрагменті.

Для формування мінімального значення рівня ймовірності $G$ (1, 2) було проаналізовано 10 текстів із наукової конференції (Технічні науки та технології – №3 (5)), які містили в собі в середньому 20000 символів. До уваги бралися терміни з найбільшими кількостями повторень в тексті. Після аналізу інтенсивності розподілу виявилося, що в чотирьох документах 7 термінів розподілені рівномірно. Приклад рівномірного розподілу терміну *tx1* на фоні нерівномірного розподілу іншого терміну *tx2* зображено на рис. 2. По осі абсцис відображені діапазони, по осі ординат відображено кількість появи терміну.

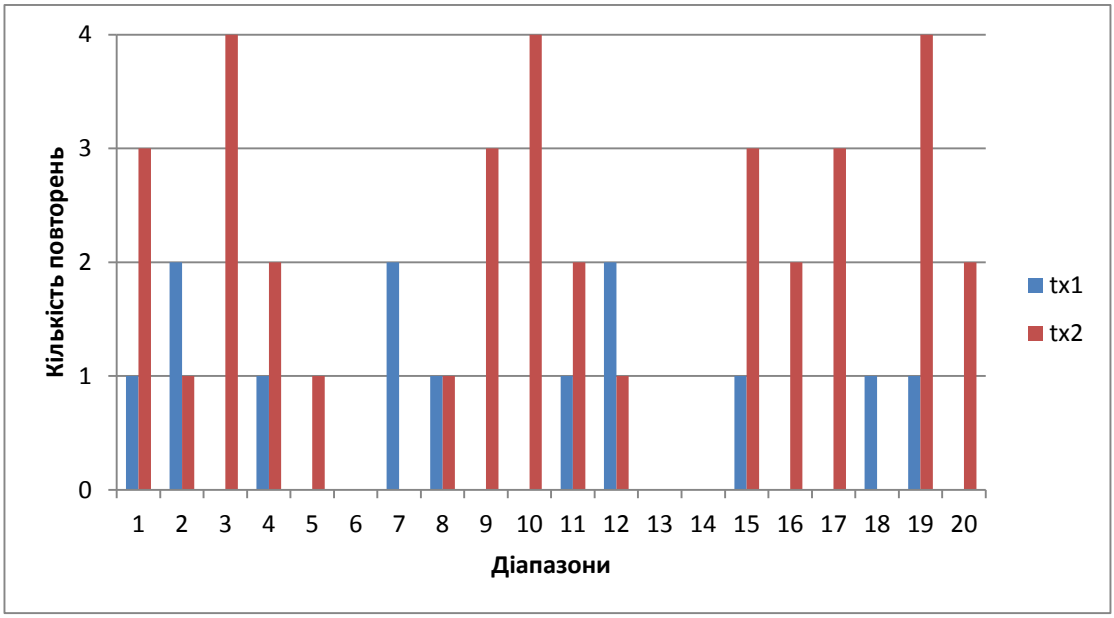


Рис. 2. Розподілення термінів по документу

Для терміну *tx1* значення рівня ймовірності $G = 0.52$, для терміну *tx2* дане значення становить $G = 1.99$. Відповідно до проведеного досліду, можна стверджувати, що терміни з значенням $G < 1$, можна рахувати рівномірно розподіленими по документу.

Визначення ефективності здійснювалося на основі формули (27) для попередніх 10 текстів. Для перших двох термінів (кожного документу), які найчастіше зустрічалися було знайдено фрагменти відповідно до першого підходу. Результати представлені в таблиці 1. Величина $Pg$ показує відсоток помилок, які виявлялися вручну, шляхом аналізу фрагментів.

Таблиця 1

Визначення ефективності використання фрагментів

| | $\sum_{z=1}^{n} txFragLen_z$ | $SCount$ | $Ef_1$ | $Pg_1$ | $\sum_{z=1}^{n} txFragLen_z$ | $SCount$ | $Ef_2$ | $Pg_2$ |
|---|---|---|---|---|---|---|---|---|
| **T1** | 143 | 257 | 44,5% | 3,3% | 177 | 257 | 31,1% | 4,1% |
| **T2** | 136 | 320 | 57,5% | 5% | 180 | 320 | 43,7% | 1,6% |
| **T3** | 200 | 280 | 28,5% | 2,8% | 148 | 280 | 47,1% | 2,1% |
| **T4** | 88 | 210 | 58% | 4% | 111 | 210 | 47,1% | 10,4% |
| **T5** | 150 | 340 | 55,8% | 9% | 210 | 340 | 38,2% | 0,8% |
| **T6** | 206 | 305 | 32,4% | 1,3% | 183 | 305 | 40% | 3% |
| **T7** | 171 | 246 | 30,5% | 2,3% | 158 | 246 | 35,8% | 2,4% |
| **T8** | 127 | 273 | 53,8% | 8,7% | 148 | 273 | 45,8% | 8,2% |
| **T9** | 175 | 295 | 40,7% | 4,5% | 163 | 295 | 44,7% | 5,7% |
| **T10** | 110 | 240 | 54,2% | 2% | 141 | 240 | 41,3% | 9,4% |

Порівняння підходів між собою здійснювалося на основі попередніх 10 текстів. До уваги бралися перші чотири терміни, які найчастіше зустрічалися і відносилися до змісту тексту.

При використанні першого підходу було виявлено наступні переваги та недоліки. *Перевагами* першого підходу є висока точність та менша кількість виконаної роботи для знаходження фрагментів одного терміну в порівнянні з наступними підходами. *Недоліками* першого підходу є неефективність порівняння фрагментів між різними термінами, так як доводиться зводити їхні діапазони до спільного розміру, що веде за собою значну кількість обрахувань. Тому даний метод не ефективний для порівняння фрагментів різних термінів і для подібних задач не буде використовуватися.

При використанні другого підходу було виявлено наступні переваги та недоліки. *Перевагами* другого підходу є наявність однакового діапазону, що значно зменшує обсяг роботи. Також перевагою є можливість порівнювати фрагменти між собою без додаткових операцій. *Недоліками* другого підходу є наявність великої кількості діапазонів, і відповідно велику кількість ітерацій роботи, через те, що прийнятий діапазон малого розміру з розрахунку на отримання точних результатів в невеликих текстах.

При використанні третього підходу було виявлено наступні переваги та недоліки. *Перевагами* третього підходу є найбільша точність формування фрагментів для кожного терміну. Наявність однакового діапазону зменшує обсяг роботи по формуванню діапазонів (як це потрібно робити в підході 1). Перевагою є можливість порівнювати фрагменти між собою без додаткових операцій. *Недоліками* третього підходу є наявність великої кількості діапазонів. Кількість усіх діапазонів становить кількості речень в тексті. Відповідно доводиться здійснювати велику кількість ітерацій обрахунків.

В результаті порівняння трьох підходів між собою, можна сказати, що перши підхід є найкращим для формування фрагментів одного терміну. Другий підхід є найефективнішим для фрагментів, які включають в себе декілька термінів. Третій підхід можна використовувати у двох випадках, при цьому точність буде найвищою, але кількість ітерацій обрахунків буде значно вищою в порівнянні в першими двома підходами.

**Висновки і пропозиції**

Розроблений метод формування фрагментів тексту для кожного терміну дозволяє зменшити час на виконання запиту користувача. Представлено алгоритм формування фрагментів. Описано принцип вибору діапазону для кожного терміну. Представлено алгоритм формування фрагментів для конкретного терміну. Також описано принцип порівняння декількох фрагментів між собою. На основі дослідження можна стверджувати, що виділення фрагментів є ефективнішим способом надання інформації в порівнянні з результатом у вигляді усього тексту. Середня ефективність виділення фрагментів для кожного терміну склала 43,5%, при цьому відсоток помилок становить 4,5%.

**Список використаних джерел**

1. Бісікало, О. В. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів [Текст] / О. В. Бісікало, В. А. Висоцька // Радіоелектроніка, інформатика, управління – №1(36), 2016. – С. 74–83.

2. Кунгурцев, О. Б. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою [Текст] / О. Б. Кунгурцев, С. В. Ковальчук, Я. В. Поточняк, М. В. Широкоступ // Технічні науки та технології – №3 (5), 2016. – С. 164–174.

3. Большакова, Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика [Текст] / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М.: МИЭМ, 2011. – 272 с.

4. Язык поисковых запросов как естественный язык [Електронний ресурс]. - Режим доступу: URL https://events.yandex.ru/lib/talks/809/.

5. Кунгурцев, О. Б. Інтерфейс для спілкування користувачів з інформаційними системами на природній мові [Текст] / О. Б. Кунгурцев, Я. В. Поточняк // Електротехнічні та комп'ютерні системи № 13 (89), 2014 – Інформаційні системи та технології.

6. Ключевые слова в тексте – как правильно употреблять? [Електронний ресурс]. - Режим доступу: URL http://dimokfm.ru/klyuchevyie-slova/.

7. Как собирать ключевые слова [Електронний ресурс]. - Режим доступу: URL http://goodseo.su/kak-sobirat-kluchevye-slova.

8. Информационные технологии управления. Методы поиска текстовой информации [Електронний ресурс]. - Режим доступу: URL https://refdb.ru/look/2575304-p10.html.

9. Распределение ключевых слов по страницам сайта [Електронний ресурс]. - Режим доступу: URL http://www.kadrof.ru/mk/4815.

10. Вавіленкова, А. І. Логіко-лінгвістичні моделі речень як засіб порівняння текстових документів за змістом [Текст] / А. І. Вавіленкова //

Математичні машини і системи. – 2012. – №1. – С. 166–173.

11. Вавіленкова, А. І. Структура лінгвістичного процесу системи порівняльного аналізу текстів за змістом [Текст] / А. І. Вавіленкова // Системи підтримки прийняття рішень. Теорія і практика: зб. доп. наук.-практ. конф. з міжнар. участю. – Київ: ІПММС НАНУ, 2011. – С. 153–156.

12. Индекс удобочитаемости [Електронний ресурс]. - Режим доступу: URL https://ru.wikipedia.org/wiki.

13. LanguageTool [Електронний ресурс]. – Режим доступу: URL https://languagetool.org/uk/.

14. Кунгурцев, А. Б. Учет межфразовых связей при автоматизированном построении толкового словаря предметной области [Текст] / А. Б. Кунгурцев, А. И. Гаврилова, А. С. Леонгард, Я. В. Поточняк // Информатика и математические методы в моделировании - №2, 2016 С.173–183.

15. Вавіленкова, А. І. Способи виявлення логічних зв'язків між частинами текстових документів [Текст] / А. І. Вавіленкова // Вісник Національного технічного університету «Харківський політехнічний університет»: зб. наук. праць. – (Серія «Нові рішення в сучасних технологіях»). – 2016. – № 12 (1184). – С. 101–105.

**References**

1. Bisikalo, O. V., Vysotska, V. A. (2016). Identifying keywords on the basis of content monitoring method in Ukrainian texts [Vyyavlennya klyuchovykh sliv na osnovi metodu kontent-monitorynhu ukrayinomovnykh tekstiv]. *Radioelektronika, informatyka, upravlinnya – Radio electronics, computer science, control*, no. 1(36), pp. 74–83 (in Ukrainian).

2. Kungurtsev, A. B., Kovalchuk, S. V., Potochniak, I. V., Shirokostup, M. V. (2016). Creating the domain vocabulary on basis automated analysis of Ukrainian texts [Pobudova slovnyka predmetnoyi oblasti na osnovi avtomatyzovanogo analizu tekstiv ukrayinskoyu movoyu]. *Texnichni nauky ta texnologiyi – Technical sciences and technology*, no. 3(5), pp. 164–174 (in Ukrainian).

3. Bol'shakova, Ye. I., Klyshinskiy, E. S., Lande, D. V., Noskov, A. A., Peskova, O. V., Yagunova, Ye. V. (2011). Automatic processing of natural language texts and computational linguistics [Avtomaticheskaya obrabotka tekstov na yestestvennom yazyke i komp'yuternaya lingvistika]. M.: MI – EM, – 272 p. (In Russian).

4. The language of search queries as a natural language [IAzyk poiskovykh zaprosov kak estestvennyi iazyk]. Retrieved from: https://events.yandex.ru/lib/talks/809/.

5. Kungurtsev, A. B., Potochniak, I. B. (2014). User interface for communication users with information systems in a natural language [Interfeys dlya spilkuvannya korystuvachiv z informatsiynymy systemamy na pryrodniy movi]. *Informatsiyni systemy ta tekhnolohiyi*. *Elektrotekhnichni ta komp"yuterni systemy – Information systems and technologies. Electrical engineering and computer systems*, no. 13(89), pp. 90–99 (in Ukrainain).

6. Keywords in the text - how to use it correctly? [Kliuchevye slova v tekste kak pravilno upotrebliat?]. Retrieved from: http://dimokfm.ru/klyuchevyie-slova/.

7. How to collect keywords [Kak sobirat kliuchevye slova]. Retrieved from: http://goodseo.su/kak-sobirat-kluchevye-slova.

8. Information technology management. Methods of searching for textual information [Informatsionnye tekhnologii upravleniia Metody poiska tekstovoi informatsii]. Retrieved from: https://refdb.ru/look/2575304-p10.html.

9. Distribution of keywords by site pages [Raspredelenie kliuchevykh slov po stranitsam saita]. Retrieved from: http://www.kadrof.ru/mk/4815.

10. Vavilenkova, A. I. (2012). Logical-linguistic model clauses as a means of comparing the content of text documents [Lohiko-linhvistychni modeli rechen yak zasib porivniannia tekstovykh dokumentiv za zmistom]. *Matematychni mashyny i systemy - Mathematical Machines and Systems*, no. 1, pp. 166–173 (in Ukrainain).

11. Vavilenkova, A. I. (2011). The structure of linguistic processes of comparative analysis of texts in content [Struktura linhvistychnoho protsesu systemy porivnialnoho analizu tekstiv za zmistom]. *Systemy pidtrymky pryiniattia rishen. Teoriia i praktyka - Decision Support Systems. Theory and practice*, pp. 153–156 (in Ukrainain).

12. Flesch–Kincaid readability tests [Indeks udobochitaemosti]. Retrieved from: https://ru.wikipedia.org/wiki.

13. LanguageTool. Retrieved from: https://languagetool.org/uk/.

14. Kungurtsev, A. B., Gavrilova, A. I., Leongard, A. S., Potochniak, I. V. (2016). Accounting of inter-phrase communication for automated construction the explanatory dictionary of domain knowledge [Uchet mezhfrazovykh sviazei pri avtomatizirovannom postroenii tolkovogo slovaria predmetnoi oblasti]. *Informatika i matematicheskie metody v modelirovanii – Informatics and mathematical methods in simulation*, no. 2, pp. 173–183 (in Ukrainian).

15. Vavilenkova, A. I. (2016). Methods for identifying logical connections between parts of text

documents [Sposoby vyiavlennia lohichnykh zviazkiv mizh chastynamy tekstovykh dokumentiv]. *Visnyk Natsionalnoho tekhnichnoho universytetu «Kharkivskyi politekhnichnyi universytet» - The Bulletins of NTU «KhPI»,* no. 12 (1184), pp. 101–105 (in Ukrainian).

## FORMATION METHOD OF TEXT FRAGMENTS FOR SEARCH SYSTEMS BASED ON THE DISTRIBUTION OF TERMS BY DOCUMENT

**A. B. Kungurtsev, S. V. Kovalchuk**
*Odessa National Polytechnic University*

***Abstract.*** *Increasing the amount of information needed for people activities in various fields, growing very fast. Finding relevant information for the user is the primary goal of any information system. On the basis scientific papers of general statements text characterized only by the number of keywords in the text. It was impossible to select a fragment of text with a term that is often found. Despite of the fact that a lot of research has been done on this topic, they do not completely solve the problem of formation of text fragments. Aim of the paper is the develop the method of formation of text fragments based on the distribution of terms by document. Developed the method of forming text fragments according to the terms of the distribution of the document. Described three approaches installation the ranges and formation of text fragments. Developed the algorithm of fragments formation. Described the principle of selection range for each term. Developed the algorithm of fragments formation for a specific term. Also described the principle of comparing several fragments together. Formation method of text fragments for each term can reduce the time for user query. Based on the study can be argued that the finding fragments are an effective way to provide information in comparison with the result as the entire text.*

***Keywords:*** *range, fragment, term, natural language, information system, subject area.*

## МЕТОД ФОРМИРОВАНИЯ ФРАГМЕНТОВ ТЕКСТА ДЛЯ ПОИСКОВЫХ СИСТЕМ НА ОСНОВЕ РАСПРЕДЕЛЕНИЯ ТЕРМИНОВ ПО ДОКУМЕНТУ

**А. Б. Кунгурцев, С. В. Ковальчук**
*Одесский национальный политехнический университет*

***Аннотация.*** *Разработан метод, позволяющий выделить фрагменты текста на основе анализа распределения терминов в ответ на запрос пользователя. Описаны три подхода установления диапазона и формирования фрагментов. Представлен алгоритм формирования фрагментов. Описан принцип выбора диапазона для каждого термина. Представлен алгоритм формирования фрагментов для указаного термина. Также описан принцип сравнения нескольких фрагментов между собой.*

***Ключевые слова:*** *диапазон, фрагмент, термин, естественный язык, информационная система, предметная область.*

Отримано 16.06.2017

**Кунгурцев Олексій Борисович** – кандидат технічних наук, професор кафедри системного програмного забезпечення, Одеський національний політехнічний університет (пр. Шевченка, 1, м. Одеса, 65044, Україна). e-mail**:** abkun@te.net.ua, тел. +38-096-976-09-27.

**Kungurtsev Alexei** – PhD in Technical Sciences, Professor of Department of System Software, Odessa National Polytechnic University (Str. Shevchenko avenue, 1, Odessa, 65044, Ukraine).
**ORCID ID:** 0000-0002-3207-7315

**Ковальчук Сергій Вікторович** – аспірант кафедри системного програмного забезпечення, Одеський національний політехнічний університет (пр. Шевченка, 1, м. Одеса, 65044, Україна). e-mail**:** serhiy_kovalchuk@mail.ua, тел. +38-067-103-99-60.

**Kovalchuk Serhiy** – PhD student of Department of System Software, Odessa National Polytechnic University (Str. Shevchenko avenue, 1, Odessa, 65044, Ukraine).
**ORCID ID:** 0000-0001-7253-0631