УДК 681.3

С.О. Сідченко, С.В. Залкін, В.В. Бєлімов

*Харківський університет Повітряних Сил ім. І. Кожедуба*

# МЕТОДИКА КОМПЛЕКСНОГО АНАЛІЗУ ДОКУМЕНТУ

*Запропоновано методику комплексного аналізу документу, яка дозволяє виділити в тексті документу закінчені відрізки тексту, що відповідають визначеним тематикам і виділити з них ключові компоненти з виділенням сугестивної спрямованості тексту.*

***аналіз документу, ключові слова, контент-аналіз***

## Вступ

**Постановка проблеми.** Методи пошуку, аналізу, складання і формування текстів (інформації) представлені дуже широко. Однак системи, що реалізують ці методи, не завжди придатні для ведення інформаційно-психологічного протиборства. Більш реалізованих систем, які застосовується для аналізу і формування текстів із заданими параметрами впливу, орієнтовані на російськомовні тексти і є вузькоспрямовані. Ще одним недоліком більшості цих систем є відсутність опису їхньої математичної бази і заборона урядів більшості держав на експорт даних продуктів у повному обсязі.

Тому необхідно створювати власні інформаційно-аналітичні системи. Будь-яка система хороша тільки в тому випадку, коли хороше математично-програмно-апаратне забезпечення, з якого вона складається, і персоналу, що його обслуговує.

**Аналіз останніх досліджень і публікацій.** Існує ряд підходів до проблеми аналізу тексту документу або бібліотеці документів [1 – 4]. В них розглянуті підходи до реалізації процесу автоматизованого квазіреферування електронних документів та формування множини їх ключових слів в інформаційно-пошукових системах, які базуються на застосуванні аналізу семантичної структури тексту та його логічної сегментації.

Каталог програм та ресурсів аналізу та лінгвістичної обробки текстів в мережі Internet наведено у [5]. Найбільш поширеними є системи ВААЛ [6] та TextAnalyst.

Система ВААЛ (http://www.vaal.ru) дозволяє прогнозувати ефект неусвідомлюваного впливу текстів на масову аудиторію, аналізувати тексти з погляду такого впливу, складати тексти з заданим вектором впливу і виявляти індивідуально-психологічні якості авторів тексту.

Система TextAnalyst (http://www.analyst.ru) дозволяє побудувати семантичну мережу понять, виділених в оброблюваному тексті, з посиланнями на контекст. Присутня функція смислового пошуку фрагментів тексту з урахуванням схованих у тексті смислових зв'язків зі словами запиту. Дозволяє аналізувати текст шляхом побудови ієрархічного дерева тем/підтем, яких торкаються в тексті. Також э можливість реферування тексту.

**Постановка завдання.** Запропонувати методику комплексного аналізу документу, яка дозволяє виділити в тексті документу закінчені відрізки тексту, що відповідають визначеним тематикам і виділити з них ключові компоненти з виділенням сугестивної спрямованості тексту.

## Виклад основного матеріалу

Документ у цілому розглядається, як послідовність слів, що можуть групуватися в речення, абзаци і розділи. Іноді можна виділити основні частини документу: заголовок (назва), анотація, аналіз літератури, мета, основна частина (що складається з розділів), список літератури й ін.

Існує ряд підходів до аналізу текстів документу.

Застосування тільки статистичного підходу для аналізу змісту документу дозволяє формувати інформацію про структуру тексту лише на основі аналізу частоти входження в нього окремих слів. При цьому “центральне” слово предметної області, що зустрічається в тексті менше визначеної кількості разів, не буде віднесено до безлічі ключових слів.

На відміну від нього семантичні методи орієнтовані на виявлення змісту (предметного змісту) тексту, його тематичної спрямованості, а також на визначення зв'язків між окремими елементами тексту і тексту в цілому. При цьому якщо в складі двох речень (абзаців) є однакові ключові слова або слова з однаковим значенням то такі речення (абзаци) будемо вважати семантично зв'язними.

Лінгвістичні підходи (засновані на синтаксичних і морфологічних методах) дозволяють привести текстові форми слів документу до словникових.

Фонетичні методи аналізу дозволяють визначити сугестивну спрямованість текстів.

Контент-аналіз дозволяє виявити частоти появи в тексті визначених характеристик, що цікавлять дослідника, та дозволяє робити деякі висновки щодо

намірів творця цього тексту або можливих реакцій адресату.

Використання сукупності перерахованих методів дозволить виділити в тексті документу закінчені відрізки тексту, що відповідають визначеним тематикам і виділити з них ключові компоненти з виділенням сугестивної спрямованості тексту в цілому і його основної частини.

Алгоритм комплексного аналізу документу наведено на рис. 1. Розглянемо загальні його етапи.

Спочатку проводиться вибір документу для аналізу та його відкриття. Документ перетворюється у зручну форму.

Проводиться статистичний аналіз документу – кількісний підрахунок слів, букв, речень, рядків, абзаців, сторінок, розділів, рисунків, таблиць, переліку посилань, тощо.

Після попереднього аналізу документу проводиться завантаження словників.

Для аналізу текстів бажано (а в деяких випадках і необхідно) використовувати набір спеціально підготовлених словників. Щоб забезпечити кращу якість аналізу, необхідно передбачити можливість самостійного настроювання їх на предметну область. Настроювання словників дозволить відфільтрувати не цікавлячу інформацію в текстах і навпаки, виділити важливу її частину.

Кожний словник повинний містити три основних розділу:

– видаляємих слів – містить слова, що просто видаляються з тексту при аналізі. В основному до слів, що видаляються, відносяться прийменники, прислівники, числівники, і деякі прикметники, що у загальному випадку не несуть корисної інформації. Крім того, їхнє видалення не порушує граматичної структури речення;

– загальновикористовуємих слів – містить слова, що не видаляються при аналізі як самостійні поняття. В основному до них відносяться слова з загальним значенням, що не представляють самостійного інтересу для аналізу матеріалу тексту, малоінформативні;

– спеціальних слів – містить слова предметної області. У цьому розділі бажано виділити слова, що становлять найбільший інтерес для користувача.

При використанні словників в результаті аналізу тексту документу відфільтруються видаляєми й загальновикористовуєми слова, а всі інші вважає спеціальними словами досліджуваного тексту, на базі яких формуються ключові поняття.

Як ключові поняття виділяються:

– спеціальні слова предметної області і їхні закінчені співвідношення, у тому числі з загальновикористовуєми словами, що зустрілися в тексті не менш заданого числа раз;

– слова-переваги, що зустрілися в тексті хоча б один раз, а також їхні закінчені сполучення, у тому числі зі спеціальними й загальновикористовуєми словам, що зустрілися в тексті не менш заданого числа раз.

Гранична частота зустрічальності, що використовувана для виділення понять, визначається в залежності від обсягу оброблюваного тексту.

Для приведення різних форм до загальної (нормалізації) може використовуватися два способи: автоматичне виділення кореневої основи і повне перерахування словоформ. У першому, більш простому випадку, досить представити в словнику одну (довільну) форму слова. У процесі аналізу всі інші форми, що мають цей же корінь, будуть вважатися еквівалентними. В другому, більш трудомісткому, але іноді необхідному випадку, усі форми слова задаються у виді списку, що будуть вважатися еквівалентними при аналізі тексту.

Потім проводиться розрахунок вагових коефіцієнтів ключових слів.

Ваговий коефіцієнт $\varpi_t$ ключового слова $t \in S$ обчислюється, як співвідношення кількості речень тексту документу, в яких зустрічається дане слово (поняття), до загальної кількості речень тексту:

$$\varpi_t = \frac{N_t}{c}, \qquad (1)$$

де $N_t$ – кількість речень тексту документу, в яких зустрічається ключове слово (поняття) $t$; $c$ – загальна кількість речень тексту.

Визначення фонетичної спрямованості ключових слів проводиться з метою визначення сугестивного навантаження ключових слів на сприйняття чоловіка.

Визначення тематики тексту. Тематика тексту документу може бути визначена шляхом порівняння ключових слів $t \in S$ тексту документу (з урахуванням їх вагових коефіцієнтів $\varpi_t$) з ключовими словами, що можуть бути віднесені до однієї з тематик.

Після вибору тематики її можливо уточнити шляхом визначення семантичного зв'язку ключових слів даної тематики, ключових слів тексту документу та назви документу (за її наявності).

Визначення сумісних компонентів сусідніх речень у тексті. Загальні елементи множин двох сусідніх речень визначаються наступних образом:

$$S_{ij} = P_i \cap P_j, \; j = i+1, \; i = \overline{1,c}, \qquad (2)$$

де $P_i$, $P_j$ – множина спеціальних слів речень $i$, $j$ відповідно; $c$ – загальна кількість речень тексту; $S_{ij}$ – множина, що складається з однакових елементів множин $P_i$ та $P_j$.

Визначення семантичного зв'язку між реченнями. Ступень семантичного зв'язку речення $i$ з наступним реченням $j$ рівняється кількості елементів множини $S_{ij}$.

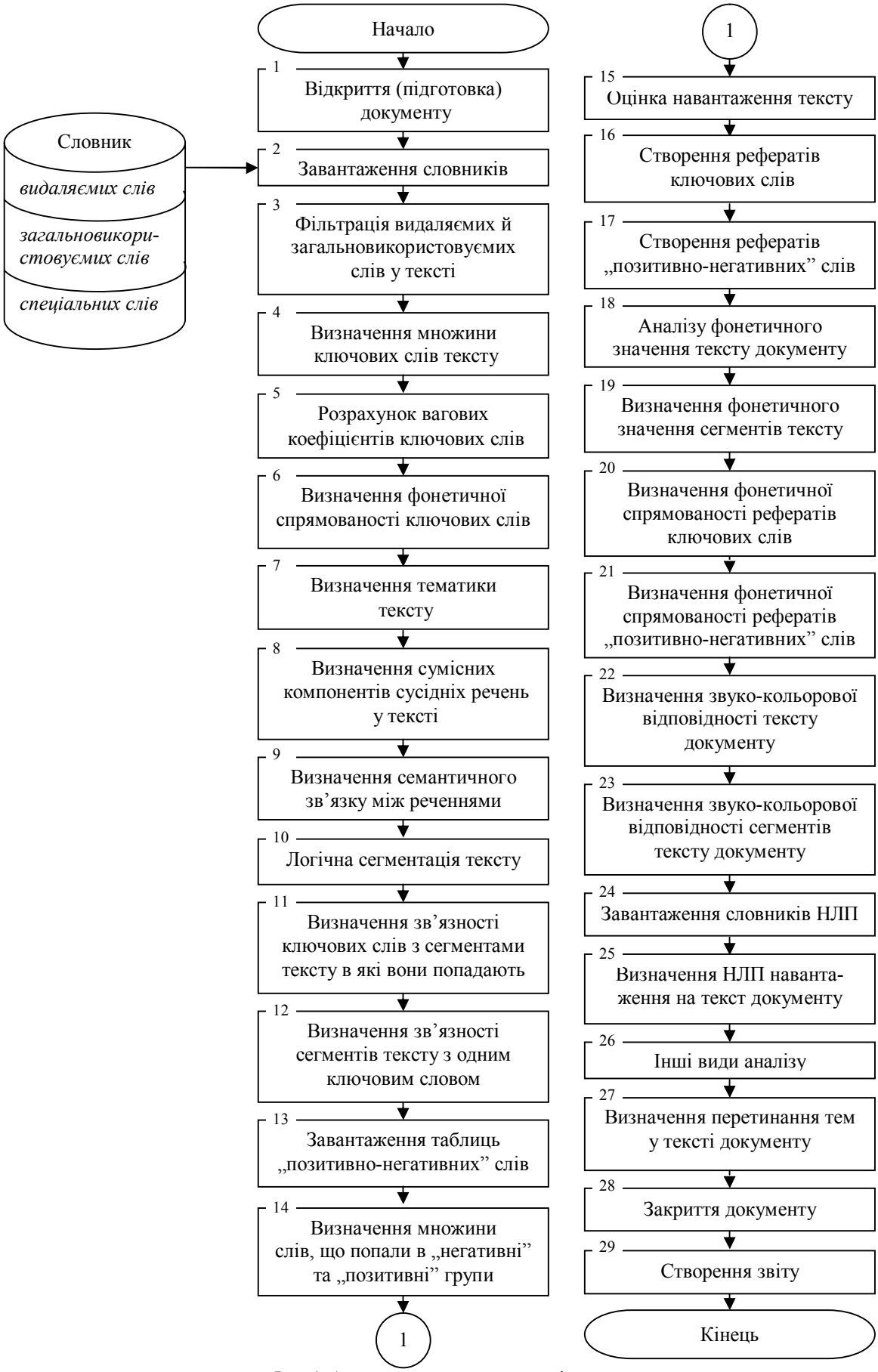


Рис. 1. Алгоритм комплексного аналізу документу

Логічна сегментація тексту. Під сегментом тексту розуміється логічно неперетинаючи відрізки тексту (речення, абзаци або їх множина), в кожному з яких мова їде о будь-якої характерної для нього інформації.

Для визначення ступеня „насиченості" логічних сегментів тексту інформацією, яка характерна для них, введемо лінійний коефіцієнт $U_i$ – параметр, який характеризує середню ступень семантичної зв'язності і-го речення з $n$ реченнями, що знаходяться попереду:

$$U_i = \frac{1}{n}\sum_{w=1}^{n}\left|S_{i-w,j-w}\right|,\ j = i+1,\ i = \overline{1,c}, \qquad (3)$$

де $n$ – кількість речень відрізку тексту, що передують і-му реченню; $\left|S_{i-w,j-w}\right|$ – ступень семантичного зв'язку речення $i-w$ з реченням $j-w$, $w = \overline{1,c}$.

Початком нового сегменту $f_k \in F$ в тексті, а відповідно, і границею між сегментами будемо вважати зменшення лінійного коефіцієнту $U_i$ поточного речення $i$, в порівнянні з лінійним коефіцієнтом $U_{i-1}$ попереднього речення $i-1$, за умови, що між поточним реченням $i$ та наступним реченням $j$ семантичний зв'язок відсутній:

$$f_k : U_i < U_{i-1},\ \left|S_{ij}\right| = 0. \qquad (4)$$

Інакше сегмент продовжується до речення, в якому семантичний зв'язок с попереднім реченням відсутній.

Визначення зв'язності ключових слів з сегментами тексту в які вони попадають визначається розрахунком семантичної ваги $L_t$ слова $t \in S_k$, яка характеризує кількість речень сегментів семантично зв'язаних словом $t$:

$$L_t = \sum_{i=1}^{t}\left|P_i \cap t\right| - 1,\ t \in S_k, \qquad (5)$$

де $t$ – елемент із множини $S_k$ ключових слів тексту документу.

Визначення зв'язності сегментів тексту з одним ключовим словом.

Ступень семантичного зв'язку сегментів тексту з одним ключовим словом рівняється кількості елементів множини $S_k$ загальних компонентів (слів) множин спеціальних слів сегментів $F_k \in F$ з одним ключовим словом, яка визначається наступним образом:

$$S_k = F_i \cap ... \cap F_j,\ i \neq j,\ k = \overline{1,l}, \qquad (6)$$

де $F_i$, $F_j$ – множина спеціальних слів сегментів $i$, $j$ відповідно, що містять однакове ключове слово; $l$ – кількість речень усіх сегментів з одним ключовим словом; $S_k$ – множина, що складається з однакових елементів множин $F_i$ та $F_j$.

Завантаження таблиць „позитивно-негативних" слів. Таблиці „позитивно-негативних" слів застосовуються для визначення ділянок тексту з однаковим набором слів, що відносяться до однієї групи термінів. Назва таблиць є умовною і може бути змінена.

Визначення множини слів, що попали в „негативні" та „позитивні" групи. Визначається множина $Y$, що складається зі слів, яки попали в „позитивні" групи, та множина $Z$ – зі слів, яки попали в „негативні" групи.

Оцінка навантаження тексту визначається, як співвідношення кількості речень тексту документу, в яких зустрічаються слова з „негативної" („позитивної") множини, до загальної кількості речень тексту:

$$v = \frac{H}{c}, \qquad (7)$$

де $H$ – кількість речень тексту документу, в яких зустрічається слова з „позитивної" $Y$ („негативної" $Z$) множини; $c$ – загальної кількості речень тексту.

Створення рефератів ключових слів. Виконується для кожного ключового слова (поняття) шляхом вибору з тексту документу усіх речень, в яких воно зустрічається.

В деяких випадках до реферату можуть включатись і цілі абзаци (сегменти) тексту, якщо це ключове слово є зв'язним.

Створення рефератів „позитивно-негативних" слів. Виконується на основі вибору речень (абзаців, сегментів) тексту документу, що містять у собі однакові ключові компоненти множин „позитивних" („негативних") слів відповідних таблиць (словників).

Аналізу фонетичного значення тексту документу. Фонетичний аналіз текстів заснований на аналізі буквеного запису з урахуванням м'якості приголосних. Текст як би представлений у звукобуквеному виді.

Визначення фонетичного значення сегментів тексту проводиться з метою простеження динаміки зміні сугестивного навантаження на сприйняття чоловіка та визначення фонетичного навантаження на різні ділянки тексту.

Воно проводиться в отриманих шляхом логічної сегментації відрізках тексту або шляхом розбивки тексту документу на рівні відрізки (як правило не менше 10).

Визначення фонетичної спрямованості рефератів ключових слів проводиться з метою простеження сугестивного навантаження на сприйняття чоловіка фрагментів тексту з однаковими ключовими словами.

Визначення фонетичної спрямованості рефератів „позитивно-негативних" слів проводиться з метою простеження сугестивного навантаження на сприйняття чоловіка фрагментів текст, що побудовані з використанням слів з „позитивних" та „негативних" груп.

Визначення звуко-кольорової відповідності тексту документу. Своєрідний аспект символіки звуків мови являє собою звуко-кольорові відповідності. Вони мають чисто сінестетичну основу і стосуються лише деяких звуків.

Найбільш чіткі звуко-кольорові відповідності простежуються в основному для голосних звуків російської мови: э, о, ы, у, и, а.

Видно, що найбільше виразно в чисті кольори пофарбовані три звуки: а – яскраво-червоний, и – голубий, о – ясно-жовтий. Помітимо, що ці три кольори є в спектрі основними в тім змісті, що шляхом змішання їх у різних пропорціях можна одержати всі інші кольори і відтінки.

Визначення звуко-кольорової відповідності тексту документу проводися шляхом підрахунку гласних букв. Текст документу приймає фарбування домінуючих гласних букв. Домінування визначається перевищенням кількості голосних букв над граничним значенням. Як правило, граничне значення визначається, як 80 – 90% від максимального.

Визначення звуко-кольорової відповідності сегментів тексту документу проводиться з метою простеження динаміки зміні фарбування тексту та виявлення звуко-кольорового навантаження на різні ділянки тексту. Воно проводиться в отриманих шляхом логічної сегментації відрізках тексту або шляхом розбивки тексту документу на рівні відрізки (як правило не менше 10).

Завантаження словників нейрон-лінгвістичного програмування (НЛП). З метою визначення НЛП спрямованості (навантаження) тексту розроблені (або розробляються) словники слів, яки характерні для людей з різноманітно розвинутими органами сприйняття. В якості таких словників виступають словники слів для аудіалів, візуалів, кінистетіків.

Визначення НЛП навантаження на текст документу проводиться шляхом підрахунку кількості слів, що попадають у різні групи. Вид НЛП навантаження вибирається відповідно групі в яку попала максимальна кількість слів.

Крім того для визначення візуального навантаження проводиться підрахунок кількості графічних зображень (рисунків). Їх якість оцінюється візуально або за допомогою спеціальних методів.

Інші види аналізу. Наприклад, проведення мотиваційного аналізу "Стремление К" – "Уход От".

Визначення перетинання тем у тексті документу проводиться шляхом визначення можливої додаткової тематики без урахування ключових слів, що відносяться тільки для основної тематики документу, що була визначена попереду.

Після завершення аналізу тексту документу проводиться приведення його в первинний вигляд або закриття.

Результати аналізу заносяться в базу даних та можуть бути використані для подальшої роботи і/або коректування середніх (нормальних) оцінок ознакових шкал та словників для всіх видів аналізу.

Звіт створюється за всіма видами аналізу в графічному та текстовому вигляді.

## Висновки

Запропоновано методику комплексного аналізу документу, яка дозволяє виділити в тексті документу закінчені відрізки тексту, що відповідають визначеним тематикам і виділити з них ключові компоненти з виділенням сугестивної спрямованості тексту. За своїми можливостями відповідає сучасним вимогам корпоративного електронного документообігу і може знайти широке застосування у різних сферах діяльності.

## Список літератури

*1. Герасимов Б.М., Сергеев О.Ю., Субач И.Ю. Извлечение информационных фраз из первичных электронных документов в информационно-поисковых системах // УСиМ. – 2006. – № 1. – С. 26-29.*

*2. Рыбаков Ф.И., Руднев Е.А., Петухов В.А. Автоматическое индексирование на естественном языке. – М.: Энергия, 1980. – 160 с.*

*3. Скороходько Е.Ф. Лінгвістичні основи автоматизації інформаційного пошуку. – К.: Вища школа, 1970. – 242 с.*

*4. Сэлтон Г.А. Автоматическая обработка, хранение и поиск информации. – М.: Сов. радио, 1973. – 560 с.*

*5. Программы анализа и лингвистической обработки текстов [Електронний ресурс]. – Режим доступу: http://www.rvb.ru/soft/catalogue/index.html*

*6. Конкурентная разведка в Интернет / В.В. Дудихин, О.В. Дудихина. – М.: ООО «Издательство АСТ», Издательство «НТ Пресс», 2004. – 229 с.*

*Надійшла до редколегії 6.11.2007*

**Рецензент:** д-р техн. наук, проф. Ю.В. Стасєв, Харківський університет Повітряних Сил ім. І. Кожедуба, Харків.