УДК 004.424.6:004.6](045)
Савченко Д. С., асп.

ПЕРЦЕПТИВНА МОДИФІКАЦІЯ МЕТОДА *N*-ГРАМ ДЛЯ ОЦІНЮВАННЯ СХОЖОСТІ ПОСЛІДОВНОСТЕЙ СИМВОЛІВ ЗА КОЕФІЦІЄНТОМ СЕРЕНСЕНА

Savchenko D. S. Perceptual modification of *N*-grams method for evaluating the similarity of symbols sequences at a Sörensen index. The article states a modification of *N*-grams method using Sörensen index for quantitative evaluating the similarity of sequences of symbols that conforms the existing criteria and allows evaluating the similarity of sequences of symbols with the peculiarities of their perception by person-expert (perceptivity principle). For this purpose, corresponding rules (of criteria) calculating the distance between sequences of symbols were defined. A proposed modification of *N*-grams method includes limitations of *N*-gram with maximum length by bigram, accounting of empty symbols in the sequence, as well as establishment of different relative weight for monograms and bigrams. The proposed method allows making more effective mechanisms of automated detection and errors correction in unstructured texts.
Keywords: symbols sequence, Sörensen index, *N*-grams method, perceptivity principle, distance between symbols sequences, similarity of symbols sequences

Савченко Д. С. Перцептивна модифікація метода *N*-грам для оцінювання схожості послідовностей символів за коефіцієнтом Серенсена. У статті запропоновано модифікацію методу N-грам з використанням коефіцієнту Серенсена для кількісної оцінки схожості послідовностей символів, що відповідає сформованому критерію і дозволяє оцінювати схожість послідовностей символів з урахуванням особливості їх сприйняття людиною-експертом (принципу перцептивності). Запропонований метод дозволяє будувати більш ефективні механізми автоматизованого виявлення та корегування помилок у неструктурованих текстах.
Ключові слова: послідовність символів, коефіцієнтом Серенсена, метод *N*-грам, принцип перцептивності, дистанція між послідовностями символів, схожість послідовностей

Савченко Д. С. Перцептивная модификация метода *N*-грамм для оценки схожести последовательностей символов по коэффициенту Серенсена. В статье предложено модификацию метода *N*-грамм с использованием коэффициента Серенсена для количественной оценки схожести последовательностей символов, соответствующую сформированному критерию и позволяющую оценивать схожесть последовательностей символов с учетом особенности их восприятия человеком-экспертом (принципа перцептивности). Предложенный метод позволяет строить более эффективные механизмы автоматизированного выявления и коррекции ошибок в неструктурированных текстах.
Ключевые слова: последовательность символов, коэффициент Серенсена, метод *N*-грамм, принцип перцептивности, схожесть последовательностей

Вступ та постановка задачі. Аналіз літературних джерел. В умовах швидкого зростання обсягів неструктурованої текстової інформації в сучасних інформаційно-телекомунікаційних системах актуальності набуває задача інтелектуалізації методів автоматизованої обробки неструктурованих текстів з урахуванням наявності в них випадкових помилок або умисних відхилень від норм лінгвоутворення окремих слів. Будь-яка сучасна інтелектуальна автоматизована система обробки неструктурованих текстів повинна бути здатною виявляти і виправляти такі помилки або відхилення, найпоширенішими з яких є помилки, пов’язані з неправильним написанням слів.

Задача виявлення та виправлення зазначених помилок традиційно вирішується через оцінювання схожості (або розбіжності) між кожним словом тексту, що представляється послідовністю символів певного алфавіту, та записами у словнику автоматизованої системи.

Завдання пошуку на підставі схожості з ключовою послідовністю (Approximate Search) розв’язуються в рамках теорії інформаційного пошуку як складової теорії інформації. Теоретичні засади інформаційного пошуку були закладені, починаючи з середини минулого століття, багатьма науковцями, серед яких Блейхут Р. (Blahut R.) [1], Гасфілд Д. (Gusfield D.), Наваро Г. (Navarro G.) [4, 5, 6], Селтон Г. (Salton G.), Укконен Е. (Ukkonen E.) [8] та інші.

Для оцінювання схожості послідовностей символів використовуються різноманітні метрики та дистанції, які пропонували Левенштейн В. І. [2], Дамерау Ф. (Damerau F.) [3], Хемінг (Hamming), Вагнер Р. (Wagner R. A.), Фішер М. (Fischer M. J.) [9], Манбер У.

(Manber U.), Хайнц Н. (Heintze N.) та інші. Також на практиці використовується низка коефіцієнтів схожості: Серенсена (Sörensen) [7], Жаккара (Jaccard), Кульчинського (Kulczinsky), Отіаі (Ochiai), Шимкевича-Симпсона (Szymkiewicz, Simpson) тощо.

Нині спостерігається збільшення інтересу до одержання за допомогою автоматизованих комп'ютерних систем таких рішень з виправлення помилок, що були б максимально наближеними до рішень, одержаних за допомогою експертів. Незважаючи на розробленість наукових основ відповідної проблематики та наявність пропозицій на ринку програмних продуктів, наукове супроводження цього сегменту залишається недостатнім.

Метою цієї публікації є формування правила (критерію) обчислення дистанції між послідовностями символів, яке б враховувало особливості їх сприйняття людиною-експертом, а також вибір методу обчислення цього критерію.

Постановка задачі. Визначення. Людина співвідносить окрему (відірвану від контексту) послідовність символів з одним із відомих слів, комбінацією слів або частиною більшого слова на підставі аналізу збігу їх символьно-позиційних характеристик. Тобто, чим більше однакових символів знаходяться в однакових відносних позиціях (відносно інших символів у послідовності), тим більше схожі послідовності між собою з точки зору сприйняття їх людиною.

Розглянемо послідовність символів $\overline{X}$ довжиною n у вигляді $\overline{X} = x_1 x_2 \ldots x_n$, де x_i ($i = 1, 2, \ldots, n$) – символ, який належить послідовності $x_i \in \overline{X}$, і знаходиться в позиції i. Кожний символ послідовності визначений над алфавітом $A = \{a_1, a_2, \ldots, a_k\}$ і може повторюватися в ній без всяких обмежень.

На підставі послідовності $\overline{X}$ утворимо нову послідовність $\overline{X}'$ шляхом додавання, заміни або видалення в ній одного символу, або шляхом взаємної зміни позицій будь-якої пари її символів. Як показано у [2, 3], використовуючи лише ці елементарні операції (додавання, заміни, видалення символу або зміни позицій символів), будь-яку вихідну послідовність символів можна перетворити за певну кількість кроків в будь-яку іншу цільову послідовність. Мінімальна кількість елементарних операцій, необхідних для такого перетворення, відома як дистанція Дамерау-Левенштейна між двома цими послідовностями.

Позначимо через σ кількісну оцінку схожості послідовностей символів $\overline{X}$ і $\overline{X}'$, припустимо, що $0 \le \sigma \le 1$, де нуль відповідає абсолютній несхожості (розбіжності), а 1 відповідає тотожності, повному збігу послідовностей.

Введемо умовні позначення для кількісних оцінок схожості послідовностей символів, що відповідають виконанню елементарних операцій щодо послідовності символів.

Позначимо через σ_{IM} оцінку схожості послідовностей $\overline{X}$ і $\overline{X}'$ для випадку, якщо $\overline{X}'$ утворена в результаті додавання одного будь-якого символу x_i з алфавіту A в будь-яку позицію в середині послідовності $\overline{X}$ (крім першої і останньої позиції). Іншими словами, це схожість, яка утворюється після виконання однієї елементарної операції вставки символу в середину послідовності. Позначимо через σ_{IE} схожість між послідовностями $\overline{X}$ і $\overline{X}'$ для випадку, якщо символ вставлений на першу або на останню позицію послідовності.

Позначимо через σ_{DM} оцінку схожості послідовностей $\overline{X}$ і $\overline{X}'$ після видалення будь-якого одного символу з середини $\overline{X}$, а через σ_{DE} – оцінку схожості після видалення першого або останнього символу з $\overline{X}$.

Позначимо через σ_{CM} оцінку схожості $\overline{X}$ і $\overline{X}'$ після заміни одного символу в середині послідовності $\overline{X}$ на інший символ з алфавіту A, а через σ_{CE} – оцінку схожості $\overline{X}$ і $\overline{X}'$ після заміни на інший символ з алфавіту A першого або останнього символу в $\overline{X}$.

Нарешті позначимо через σ_{TM} оцінку схожості після взаємної зміни позицій (транспозиції) двох символів, якщо обидва вони знаходяться в середині послідовності $\overline{X}$,

через σ_{TE} – оцінку схожості після взаємної зміни позицій двох символів, якщо один з них знаходиться у послідовності $\overline{X}$ на першій або на останній позиції, а другий – в середині послідовності, та через σ_{TEE} – оцінку схожості після взаємної зміни позицій першого і останнього символів.

Окреме врахування елементарних операцій за участю символів з середини послідовності і її крайніх символів пояснюється результатами досліджень у галузі когнітивної психології, які свідчать про те, що людина швидше сприймає слова з перемішаними літерами тоді, коли на своїх позиціях знаходяться перша і остання літери, а не середні (ефект Кембриджського університету).

Критерії схожості послідовностей символів. Із врахуванням зазначеного, визначимо основне правило (критерій) схожості двох послідовностей наступними двома групами умов:

$$\sigma_{IM} > \sigma_{IE},\ \sigma_{DM} > \sigma_{DE},\ \sigma_{CM} > \sigma_{CE},\ \sigma_{TM} > \sigma_{TE} > \sigma_{TEE}; \quad (1)$$

$$\sigma_{IM} > \sigma_{DM} > \sigma_{CM},\ \sigma_{IE} > \sigma_{DE} > \sigma_{CE}. \quad (2)$$

Група умов (1) встановлює те, що елементарні операції над середніми символами послідовності повинні менше впливати на схожість послідовностей $\overline{X}$ і $\overline{X}'$, ніж аналогічні операції за участю їх крайніх символів. Група умов (2) встановлює те, що елементарна операція вставки символу повинна менше впливати на схожість між послідовностями $\overline{X}$ і $\overline{X}'$ порівняно з елементарною операцією видалення символу, а елементарна операція видалення символу, в свою чергу, повинна менше впливати на схожість між послідовностями $\overline{X}$ і $\overline{X}'$ порівняно з елементарною операцією заміни символу.

Метод обчислення сформованого критерію (1-2) будемо будувати на основі методу N-грам, який (на відміну від інших) враховує символьно-позиційні характеристики послідовностей, що порівнюються.

Для цього розглянемо послідовність символів $\overline{S}_1$ у вигляді $\overline{S}_1 = x_1 x_2 \ldots x_n$, що складається із n символів x_i у позиціях $i = 1, 2, \ldots, n$, а також послідовність символів $\overline{S}_2$ у вигляді $\overline{S}_2 = y_1 y_2 \ldots y_m$, що складається із m символів y_j у позиціях $j = 1, 2, \ldots, m$. Кожний символ x_i послідовності $\overline{S}_1$ і кожний символ y_j послідовності $\overline{S}_2$ належить до множини A певного алфавіту символів $A = \{a_1, a_2, \ldots, a_k\}$, $x_i \in A$, $y_j \in A$.

Кожну із заданих послідовностей $\overline{S}_1$, $\overline{S}_2$ у загальному випадку можна уявити як сукупність окремих символів, а також – як сукупність груп сусідніх символів (N-грам): груп по два символи (біграм), груп по три символи (триграм), і т.д. аж включно до N символів (N-грам), де $N = n$ для $\overline{S}_1$, $N = m$ для $\overline{S}_2$.

Таким чином, перша послідовність $\overline{S}_1$ становить собою сукупність n символів $x_1, x_2, \ldots, x_n$, сукупність $n-1$ біграм $x_1x_2, x_2x_3, \ldots, x_{n-1}x_n$, сукупність $n-2$ триграм $x_1x_2x_3, x_2x_3x_4, \ldots, x_{n-2}x_{n-1}x_n$, і т.д. включно до 1 n-грами $x_1x_2\ldots x_n$, яка і є власне послідовністю $\overline{S}_1$.

Аналогічно друга послідовність $\overline{S}_2$ становить собою сукупність m символів $y_1, y_2, \ldots, y_m$, сукупність $m-1$ біграм $y_1y_2, y_2y_3, \ldots, y_{m-1}y_m$, сукупність $m-2$ триграм $y_1y_2y_3, y_2y_3y_4, \ldots, y_{m-2}y_{m-1}y_m$, і т.д. аж до 1 m-грами $y_1y_2\ldots y_m$, яка і є послідовністю $\overline{S}_2$.

Якщо вважати окремі символи послідовності також одним із варіантів N-грами (а саме – монограмою), то загальна кількість усіх можливих варіантів N-грам для послідовності символів довжиною n розраховується за формулою $\left(n^2 + n\right)/2$, оскільки кількості різних варіантів N-грам для цієї послідовності утворюють числовий ряд $n, n-1, n-2, \ldots, 3, 2, 1$.

Відтак, послідовність $\overline{S}_1$ має $\left(n^2+n\right)/2$ варіантів N-грам, а послідовність $\overline{S}_2$ має відповідно $\left(m^2+m\right)/2$ варіантів N-грам.

Кількісна оцінка схожості двох послідовностей символів методом N-грам полягає у зрівнянні кількості їх N-грам, що збігаються, і кількості їх N-грам, що не збігаються між собою. При цьому, як вже зазначалося, можуть бути використані різні коефіцієнти схожості.

Наприклад, оберемо для кількісної оцінки схожості двох послідовностей символів коефіцієнт схожості Серенсена, який визначається як $K=\dfrac{2c}{a+b}$, де: a – кількість елементів в першому наборі, b – кількість елементів в другому наборі, c – кількість спільних елементів для першого и другого набору, K – коефіцієнт схожості Серенсена.

Тоді схожість σ послідовностей $\overline{S}_1$ і $\overline{S}_2$ на підставі коефіцієнту схожості Серенсена можна охарактеризувати як відношення кількості їх N-грам, що співпадають між собою, до загальної кількості N-грам в обох послідовностях:

$$\sigma_{S_1S_2}=\frac{2c}{\dfrac{n^2+n}{2}+\dfrac{m^2+m}{2}}=\frac{4c}{n^2+n+m^2+m}, \tag{3}$$

де: c – кількість спільних пар N-грам для обох послідовностей, n – довжина послідовності $\overline{S}_1$, m – довжина послідовності $\overline{S}_2$.

Наприклад, розглянемо схожість послідовностей символів “ТЕКСТ” і “ТЕСТ”. Перша послідовність утворює сукупність з 15 N-грам: “Т”, “Е”, “К”, “С”, “Т”, “ТЕ”, “ЕК”, “КС”, “СТ”, “ТЕК”, “ЕКС”, “КСТ”, “ТЕКС”, “ЕКСТ” і “ТЕКСТ”. Друга послідовність утворює сукупність з 10 N-грам: “Т”, “Е”, “С”, “Т”, “ТЕ”, “ЕС”, “СТ”, “ТЕС”, “ЕСТ” і “ТЕСТ”. Як видно, 6 пар N-грам в обох послідовностях спільні: “Т”, “Е”, “С”, “Т”, “ТЕ”, “СТ”. Тоді кількісна оцінка схожості зазначених послідовностей дорівнює:

$$\sigma\left(TEKCT,TECT\right)=\frac{4\times 6}{5^2+5+4^2+4}=0{,}48.$$

Як можна побачити у випадку, коли обидві послідовності повністю збігаються, їх кількісна оцінка схожості σ завжди дорівнює 1. Коли в обох послідовностях немає навіть жодного спільного символу, їх кількісна оцінка схожості дорівнює 0. В інших випадках кількісна оцінка схожості послідовностей σ, як і потрібно, знаходиться у діапазоні (0, 1).

В той же час, подібний метод оцінки передбачає, що середня частина послідовності повинна мати більшу вагу на її схожість з іншою послідовністю, ніж початок і кінець послідовності, оскільки її середні символи входять в більшу кількість N-грам, аніж крайні символи. І ця відносна вага середніх символів у порівнянні з крайніми зростає із зростанням довжини послідовності.

Зазначене добре видно з наведеного прикладу схожості послідовностей “ТЕКСТ” і “ТЕСТ”, коли відсутність 1 середньої літери “К”, яка входить до 9 N-грам з 15, призвела до зменшення кількісного показника схожості більш, ніж вдвічі, з 1,00 до 0,48. У випадку відсутності крайнього символу замість середнього, наприклад, при порівнянні тим же методом послідовностей “ТЕКСТ” і “ЕКСТ”, їх схожість складатиме 0,80.

Варіанти методу *N*-грам. Однак на практиці, як вже зазначалося і було відображено у (1-2), спостерігається зворотна ситуація, коли крайні частини послідовності символів мають більше значення для успішного розпізнавання слова, ніж її середня частина.

Враховуючи зазначене, слід констатувати, що без модифікації, у своєму "класичному" вигляді метод *N*-грам не є придатним для кількісного оцінювання схожості послідовностей символів за визначеним критерієм (1-2). Тому виникає потреба дослідити і порівняти між собою різні його варіанти.

Зокрема, розглянемо такі варіанти методу *N*-грам:

– з використанням однієї *N*-грами фіксованої довжини;

– з використанням *N*-грам довжиною не більше встановленої (тобто, з обмеженням максимальної довжини *N*-грами);

– з використанням "пустих" символів на початку та в кінці послідовності (для доповнення до *N*-грами потрібної довжини);

– зі встановленням різної відносної ваги для *N*-грам різної довжини.

Оцінка з фіксованою довжиною *N*-грами передбачає врахування у послідовностях лише одного виду *N*-грам, наприклад, біграм або триграм тощо. При цьому, слід враховувати декілька особливостей оцінки з фіксованою довжиною *N*-грами.

По перше, така оцінка не може бути застосована для послідовностей, довжина яких менше фіксованої довжини N-грами, оскільки такі послідовності не міститимуть в собі жодної N-грами. Наприклад, неможливо без введення додаткових умов порівняти послідовності "ВИ" і "ТИ" через аналіз їх триграм, оскільки ні перша, ні друга послідовність не містять жодної триграми. Можливим варіантом подолання цього обмеження є обов'язкове доповнення послідовностей з обох сторін "пустими" символами і врахування цих "пустих" символів у N-грамах (до потрібної довжини). У такому разі послідовність "ВИ" складатиметься із триграм "¤¤В", "¤ВИ", "ВИ¤" і "И¤¤", а послідовність "ТИ" складатиметься із триграм "¤¤Т", "¤ТИ", "ТИ¤" і "И¤¤", а формула (3) матиме вигляд:

$$\sigma_{S_1S_2} = \frac{2c}{n+m+4}.$$

По-друге, у зв'язку із залежністю від довжини *N*-грами відносної ваги середніх символів послідовності у порівнянні з крайніми символами у варіантах з більшою довжиною *N*-грам вплив середніх символів на схожість послідовностей буде більшим, ніж у варіантах з меншою довжиною *N*-грам. А оскільки, відповідно до критерію (1-2), відносна вага середніх символів у послідовності повинна бути меншою, ніж вага її крайніх символів, можна зробити висновок, що варіанти методу з меншою довжиною *N*-грам більше відповідатимуть цій умові, ніж варіанти з більшою довжиною *N*-грам.

По-третє, при використанні лише монограм відсутня залежність кількісної оцінки схожості від відносних позицій, які займають символи у послідовностях. Як наслідок, послідовності з однаковим набором символів, розташованих у різному порядку (наприклад, "ВЕКТОР" і "КОРВЕТ"), стають такими, що не відрізняються (їх схожість дорівнює 1,00). Це суттєво обмежує застосування методу.

Таким чином, слід очікувати, що серед варіантів з фіксованою довжиною *N*-грами для практичної реалізації найбільш привабливим буде варіант із врахуванням біграм з доповненням послідовності пустими символами (оскільки це є варіант з найменшою довжиною *N*-грами, яка при цьому враховує порядок слідування символів).

Оцінка з обмеженням максимальної довжини N-грами передбачає врахування у послідовностях декількох видів N-грам від монограми до N-грами заданої довжини (біграми, триграми тощо). Для цієї групи варіантів характерні аналогічні особливості, як і для варіанту з фіксованою довжиною N-грами. Тому слід очікувати, що для практичного використання найбільш привабливою серед варіантів цієї групи буде варіант з обмеженням максимальної довжини N-грами до біграм (тобто, монограми та біграми) з доповненням послідовності пустими символами. При цьому формула (3) матиме вигляд:

$$\sigma_{S_1S_2} = \frac{2c}{2(n+m+1)} = \frac{c}{(n+m+1)}.$$

Для порівняння відповідності різних варіантів оцінки схожості послідовностей символів критерію (1-2) було розглянуто 9 різних варіантів методу *N*-грам на прикладі вихідних слів (послідовностей символів) "БАНК" і "ТЕРМІН" довжиною 4 і 6 символів відповідно. Результати розрахунків наведено у табл. 1 і табл. 2.

Зокрема, порівнювалися такі варіанти:

1) 3 варіанти з фіксованою довжиною *N*-грами (лише монограми – колонки I, лише біграми без пустих символів – колонки II, лише біграми з пустими символами – колонки III);

2) 4 варіанти з обмеженням максимальної довжини *N*-грами (до біграм без пустих символів – колонки IV, до біграм з пустими символами – колонки V, до триграм без пустих символів – колонки VI, до триграм з пустими символами – колонки VII);

3) 2 варіанти без обмеження максимальної довжини *N*-грами (*N*-грами без пустих символів – колонки VIII, *N*-грами з пустими символами – колонки IX).

Схожість слова, що містить помилку, з записом у словнику довжиною 4 символи за різними варіантами методу *N* -грам, з точністю до 0,01 Табл. 1

Запис у словнику	БАНК	I	II	III	IV	V	VI	VII	VIII	IX
Транспозиція середніх σ_{TM}	БНАК	1,00	0	0,40	0,57	0,67	0,44	0,53	0,40	0,45
Зайвий середній σ_{IM}	БЛАНК	0,89	0,57	0,73	0,75	0,80	0,67	0,73	0,56	0,70
Відсутність середнього σ_{DM}	БАК	0,86	0,40	0,67	0,67	0,75	0,53	0,67	0,50	0,71
Зайвий крайній σ_{IE}	БАНКА	0,89	0,86	0,73	0,88	0,80	0,86	0,73	0,80	0,70
Відсутність крайнього σ_{DE}	АНК	0,86	0,80	0,67	0,83	0,75	0,80	0,67	0,75	0,71
Заміна середнього σ_{CM}	БАРК	0,75	0,33	0,60	0,57	0,67	0,44	0,60	0,40	0,55
Заміна крайнього σ_{CE}	ТАНК	0,75	0,67	0,60	0,71	0,67	0,67	0,60	0,60	0,55
Транспозиція 1 крайнього σ_{TE}	БАКН	1,00	0,33	0,40	0,71	0,67	0,56	0,53	0,50	0,45
Транспозиція 2 крайніх σ_{TEE}	КАНБ	1,00	0,33	0,20	0,71	0,56	0,56	0,33	0,50	0,23

Схожість слова, що містить помилку, з записом у словнику довжиною 6 символів за різними варіантами методу *N* -грам, з точністю до 0,01 Табл. 2

Запис у словнику	ТЕРМІН	I	II	III	IV	V	VI	VII	VIII	IX
Транспозиція середніх σ_{TM}	ТЕМРІН	1,00	0,40	0,57	0,73	0,77	0,53	0,38	0,38	0,51
Зайвий середній σ_{IM}	ТЕРАМІН	0,92	0,73	0,80	0,83	0,86	0,73	0,80	0,49	0,60
Відсутність середнього σ_{DM}	ТЕРІН	0,91	0,67	0,77	0,80	0,83	0,67	0,77	0,50	0,70
Зайвий крайній σ_{IE}	СТЕРМІН	0,92	0,91	0,80	0,92	0,86	0,91	0,80	0,86	0,60
Відсутність крайнього σ_{DE}	ЕРМІН	0,91	0,89	0,77	0,90	0,83	0,89	0,77	0,83	0,70
Заміна середнього σ_{CM}	ТЕРУІН	0,83	0,60	0,71	0,73	0,77	0,60	0,71	0,43	0,59
Заміна крайнього σ_{CE}	ТЕРМІТ	0,83	0,80	0,71	0,82	0,77	0,80	0,71	0,71	0,59
Транспозиція 1 крайнього σ_{TE}	ТЕРМНІ	1,00	0,60	0,57	0,82	0,77	0,73	0,67	0,57	0,51
Транспозиція 2 крайніх σ_{TEE}	НЕРМІТ	1,00	0,60	0,43	0,82	0,69	0,73	0,52	0,57	0,20

В кожну із заданих послідовностей символів вносилися помилки, і оцінювалася схожість утворених послідовностей із заданими. Зокрема, перевірялися: 1) транспозиція середніх символів; 2) наявність зайвого символу в середині послідовності; 3) відсутність одного символу з середини послідовності; 4) наявність зайвого символу з краю послідовності; 5) відсутність одного символу з краю послідовності; 6) заміна одного символу з середини послідовності іншим символом; 7) заміна одного символу з краю послідовності іншим символом; 8) транспозиція двох символів, один з яких – з середини послідовності, а інший – з її краю; 9) транспозиція двох крайніх символів послідовності.

Як видно з наведених таблиць, одержані результати в цілому подібні для різних заданих послідовностей символів. Не дивлячись на те, що числові значення схожості в них різні (оскільки вони залежать від довжини послідовностей), їх відносне розташування за типами помилок в перших 5 варіантах збігається повністю, а в останніх 4 варіантах – збігається з

деякими незначними відхиленнями, що пояснюється більшою залежністю таких варіантів від довжини послідовностей.

Також видно, що критерію (1-2) варіанти методу відповідають лише частково. Основним недоліком при цьому є невідповідність групі умов (1): у кращому випадку – відсутність різниці між аналогічними операціями над середніми і крайніми символами (колонки I, III, V, VII, IX), у гіршому випадку – перевага середніх символів над крайніми (колонки II, IV, VI, VIII). Інші умови критерію (1-2) не виконуються у колонках I та IX. Отже, можна зробити висновок, що сформованому критерію (1-2) серед розглянутих варіантів найбільш відповідають варіанти із врахуванням пустих символів: 1) лише біграми (колонка III); 2) монограми і біграми (колонка V); 3) монограми, біграми і триграми (колонка VII).

Для подолання основного виявленого недоліку щодо відповідності групі умов (1) призначимо різну вагу *N*-грамам різної довжини. Як вже зазначалося, для цього не підходять варіанти із фіксованою довжиною *N*-грами. Тому, з урахуванням результатів порівняння, введемо різну вагу *N*-грам для варіанту з обмеженням максимальної довжини *N*-грами до біграми (колонка V). Зокрема, зменшимо вагу середніх біграм порівняно з монограмами і крайніми біграмами. Для цього, наприклад, встановимо, що кожна монограма, а також біграма з пустим символом (з краю послідовності) мають вагу у 2 рази більше, ніж звичайна біграма (в середині послідовності). Це виглядає доволі логічним, оскільки відсутність одного символу в середині послідовності впливає відразу на 2 біграми при кількісній оцінці схожості і лише тільки на 1 монограму.

За таких умов загальна сума ваги всіх монограм і біграм послідовності довжиною n буде розраховуватись як $3(n+1)$, а формула (3) матиме вигляд:

$$\sigma_{S_1S_2} = \frac{2W}{3(n+1)+3(m+1)} = \frac{2W}{3(n+m+2)},$$

де W – сума ваги монограм і біграм у послідовностях, які збігаються.

У табл. 3 наведено порівняння оцінок схожості послідовностей символів “БАНК” і “ТЕРМІН”, що містять різні типи помилок, на основі варіантів із врахуванням пустих символів з обмеженням максимальної довжини *N*-грами до біграм з однаковою (колонки V), а також з різною (колонки X) вагою монограм і біграм.

Як можна переконатися, результати оцінок схожості в колонці X повністю відповідають критерію (1-2).

Порівняння варіантів методу *N*-грам з однаковою (V) та різною (X) вагою на прикладі слів довжиною 4 і 6 символів, з точністю до 0,01 Табл. 3

Запис у словнику	**БАНК**	**V**	**X**	**ТЕРМІН**	**V**	**X**
Транспозиція середніх σ_{TM}	БНАК	0,67	0,80	ТЕМРІН	0,77	0,86
Зайвий середній σ_{IM}	БЛАНК	0,80	0,85	ТЕРАМІН	0,86	0,89
Відсутність середнього σ_{DM}	БАК	0,75	0,81	ТЕРІН	0,83	0,87
Зайвий крайній σ_{IE}	БАНКА	0,80	0,79	СТЕРМІН	0,86	0,84
Відсутність крайнього σ_{DE}	АНК	0,75	0,74	ЕРМІН	0,83	0,82
Заміна середнього σ_{CM}	БАРК	0,67	0,73	ТЕРУІН	0,77	0,81
Заміна крайнього σ_{CE}	ТАНК	0,67	0,67	ТЕРМІТ	0,77	0,76
Транспозиція 1 крайнього σ_{TE}	БАКН	0,67	0,73	ТЕРМНІ	0,77	0,81
Транспозиція 2 крайніх σ_{TEE}	КАНБ	0,56	0,60	НЕРМІТ	0,69	0,71

Висновки. Для автоматизованого виявлення та корегування помилок у неструктурованих текстах визначені правила (критерії) обчислення дистанції між послідовностями символів з урахуванням особливості їх сприйняття людиною-експертом

(перцептивності), де враховано те, що, елементарні операції над середніми символами послідовності повинні менше впливати на схожість послідовностей, ніж аналогічні операції за участю їх крайніх символів, а також те, що елементарна операція вставки символу повинна менше впливати на схожість між послідовностями порівняно з операцією видалення символу, а елементарна операція видалення символу повинна менше впливати на схожість між послідовностями порівняно з елементарною операцією заміни символу.

Запропонована модифікація методу N -грам з використанням коефіцієнту Серенсена для кількісної оцінки схожості послідовностей символів, що відповідає сформованому критерію і дозволяє оцінювати схожість послідовностей символів з урахуванням перцептивності. Ця модифікація включає обмеження біграмами максимальної довжини N -грами, використання пустих символів, встановлення різної відносної ваги для монограм та біграм.

Проте слід зазначити, що недоліком запропонованого методу оцінки схожості послідовностей символів в системах автоматизованої обробки неструктурованих текстів є неможливість уникнути повного перебору словника і подальшого сортування результатів для встановлення кожного разу найбільш імовірних еквівалентів словникових статей для заданого текстового фрагменту.

Література

1. Блейхут Р. Теория и практика кодов, контролирующих ошибки / Р. Блейхут ; пер. с англ. – Москва : Мир, 1986. – 576 с.

2. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов В.И. Левенштейн / Докл. Академий Наук СССР. – 1965. – т.163.4. – С. 845-848.

3. Damerau F.A. Technique for Computer Detection and Correction of Spelling Errors / F.A. Damerau // Communications of the ACM. – 1964. – Vol. 7. – No. 3. – P. 171-176.

4. Hyyrö H. Faster bit-parallel approximate string matching / H. Hyyrö, G. Navarro // Proc. 13th Combinatorial Pattern Matching (CPM'2002), LNCS 2373. – 2002. – PP. 203-224.

5. Navarro G. A guided tour to approximate string matching / G. Navarro // ACM Computing Surveys. – 2001. No. 33(1). – PP. 31-88.

6. Navarro G. A practical q-gram index for text retrieval allowing errors / G. Navarro, R. Baeza-Yates // CLEI Electronic Journal. – 1998. – No. 1(2).

7. Sörensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content / T. Sörensen // Kongelige Danske Videnskabernes Selskab. Biol. krifter. Bd V. – 1948. – № 4. – P. 1-34.

8. Ukkonen E. Approximate string-matching with q-grams and maximal matches / E. Ukkonen // Theoretical Computer Science 92. – 1992. – P. 191-211.

9. Wagner R. A. The String-to-string Correction Problem / R.A. Wagner , M.J. Fischer // Journal of ACM. – 1974. – Vol. 21. – No. 1. – P. 168-173.

Автор статті

Савченко Денис Сергійович – аспірант кафедри інформаційних систем і технологій та захисту інтересів держави у сфері інформаційної безпеки Національної академії Служби безпеки України, м. Київ. Тел.: +380 (67) 244 14 58. E-mail: sdensys@gmail.com.

Author of the article

Savchenko Denys Serghiyovych – post-graduate of department of information systems and technologies and protection of state interests in the field of information security, National Academy of Security Service of Ukraine, Kyiv. Tel.: +380 (67) 244 14 58. E-mail: sdensys@gmail.com.

Дата надходження в редакцію: 02.03.2016 р. Рецензент: д.т.н., проф. О.В. Барабаш