УДК 004.6, 004.8

КЛАСТЕРИЗАЦІЙНІ ПАРАМЕТРИ ДЛЯ ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ ПРИРОДНИХ ТЕКСТІВ ІЗ УРАХУВАННЯМ ПОПРАВОК НА РАНДОМНІ ТЕКСТИ

Олег Кушнір, Василь Васюта, Богдан Горон, Іван Довгань

CLUSTERING-BASED PARAMETERS FOR DETECTION OF KEYWORDS IN NATURAL TEXTS WITH CORRECTIONS ON RANDOM TEXTS

Oleh Kushnir, Vasyl Vasiuta, Bohdan Horon, Ivan Dovhan

кафедра оптоелектроніки та інформаційних технологій, Львівський університет імені Івана Франка, вул. Тарнавського 107, м. Львів, 79017, Україна

Abstract. *Clustering-based methods are efficient in detection of keywords in natural texts. Here we present a technique for correcting the appropriate word-relevance parameters with the data obtained for random texts and simulate the dependences of those parameters on the absolute and relative word-token frequencies.*

Визначення ключових слів і фраз текстів – один із важливих напрямів опрацювання природної мови та інформаційного пошуку і класифікації текстових документів. Попри сучасну тенденцію до пошуку ключових слів із застосуванням підходів машинного навчання, актуальним є розвиток базових, інтуїтивно зрозумілих і аналітично обґрунтованих методів і алгоритмів пошуку, заснованих на властивостях розташування токенів різних слів у досліджуваному тексті, які не вимагають референтної текстової бази.

З літератури відомий спосіб ранжування слів за релевантністю та знаходження ключових слів у тексті [1], який припускає, що повторні появи токенів семантично бідних стоп-слів у тексті описуються стохастичним процесом, який у континуальному наближенні приводить до експоненційного розподілу ймовірності $p(\tau) \sim \exp(-\tau)$ т. зв. часів очікування токенів τ, а токени ключових слів, що стисло виражають семантику тексту, виявляють явище кластеризації, тобто групуються на деяких ділянках тексту та відсутні на інших його ділянках.

Важливий наслідок кластеризації – поява важкого хвоста розподілу $p(\tau)$, а найпростіший кількісний індикатор його наявності – параметр $R = k_{21} = (\mu_2)^{1/2} / \mu_1$ [1], де μ_1 і μ_2 – перші кумулянти розподілу $p(\tau)$. Справді, якщо $p(\tau)$ експоненційний, то $(\mu_2)^{1/2} = \mu_1$ і $R = 1$. Ту ж ситуацію спостерігатимемо у рандомному тексті (RT), наприклад у рандомізованій версії вихідного природного тексту, утвореній шляхом багатократних перестановок слів у ньому. Важчий же хвіст розподілу $p(\tau)$ для деякого ключового слова в природному тексті означатиме умову $R > 1$ і кластеризацію цього слова. Нарешті, гіпотетичний граничний випадок $R = 0$ описував би періодичне розміщення токенів слова в тексті. Відповідно, ранжований за спаданням R список слів тексту даватиме деяке наближення списку його ключових слів.

Окрім параметра R, у літературі аналізували використання інших комбінацій кумулянтів на зразок коефіцієнта асиметрії $k_{32} = \mu_3 / (\mu_2)^{3/2}$ розподілу $p(\tau)$ [2], а також складнішого коефіцієнта кластеризації Γ [3]. У цій праці ми вперше перевіряли застосовність до проблеми визначення ключових слів інших можливих комбінацій кумулянтів порядків 1÷4, тобто параметрів k_{31}, k_{41}, k_{42} (коефіцієнта ексцесу) і k_{43}. За умови постулювання експоненційного розподілу $p(\tau)$ усі слова в RT, позбавленому будь-якої семантики, повинні описуватися однаковими параметрами $k_{21}^{RT} = 1$, $\Gamma^{RT} = 2e^{-2} \approx 0.27$ [2, 3], $k_{31}^{RT} = k_{32}^{RT} = 2$, $k_{41}^{RT} = k_{42}^{RT} = 9$ і $k_{43}^{RT} = 9/2^{4/3}$.

Проте всі названі підходи недосконалі: часи очікування τ появ токенів деякого слова в тексті як стохастичного процесу повторних бінарних випробувань Бернуллі не можна описувати неперервним експоненційним розподілом, хоча би через дискретний характер відповідних змінних. Наслідком цього грубого наближення можуть бути такі артефакти як залежність параметрів k_{ij} від абсолютної частоти даного слова F, його відносної частоти f ($f = F/L$) і навіть від довжини тексту L (кількості усіх слів у ньому). Це стає найбільш очевидним, коли розглянути RT. Він позбавлений будь-якої семантики, проте параметри k_{ij}^{RT} такого тексту теж будуть функціями $k_{ij}^{RT}(f,F,L)$ і тому різними для слів із різними частотами. Це не має жодного сенсу, оскільки RT як рандомна символьна послідовність не містить ключових слів.

Причинами описаних непослідовностей є континуальне наближення $p(\tau) \sim \exp(-\tau)$, що правильне лише за умов $f \to 0$, $\tau \to \infty$ і $L \to \infty$. Якщо ж брати скінченні $f = F/L$, τ і наближення $L \to \infty$ та $F \to \infty$, то приходимо до геометричного розподілу ймовірності часів очікування $p_{geo}(\tau) = f(1-f)^{\tau-1}$ [4]. Для уточнення даних про ключові слова можна перейти до параметрів релевантності $k_{n,ij} = k_{ij}/k_{ij}^{RT}$ і $\Gamma_n = \Gamma/\Gamma^{RT}$, які нормовані на відповідні параметри для RT. Знаходячи кумулянти розподілу $p_{geo}(\tau)$

$$\mu_1^{RT} = 1/f,\ \mu_2^{RT} = (1-f)/f^2,\ \mu_3^{RT} = (2-f)(1-f)/f^3,\ \mu_4^{RT} = (1-f)\left[9(1-f)+f^2\right]/f^4, \qquad (1)$$

на підставі означень k_{ij} легко одержати теоретично передбачувані параметри RT k_{ij}^{RT}:

$$k_{21}^{RT} = \sqrt{1-f},\ k_{31}^{RT} = (2-f)(1-f),\ k_{32}^{RT} = \frac{2-f}{\sqrt{1-f}},$$

$$k_{41}^{RT} = (1-f)\left[9(1-f)+f^2\right],\ k_{42}^{RT} = \frac{9(1-f)+f^2}{1-f},\ k_{43}^{RT} = \frac{9(1-f)+f^2}{(2-f)^{4/3}(1-f)^{1/3}}. \qquad (2)$$

Тоді знайдемо нормовані параметри R_n, $k_{n,32}$ і Γ_n ($\Gamma^{RT} = \frac{1}{2}h(h-1)(1-f)^h[(1-f)^{-1} - f - 1]$ і $h = \mathrm{Int}(2/f)$) [2, 4], а також решту параметрів $k_{n,ij}$.

Мета цієї роботи – це вивчення поведінки емпіричних параметрів RT $k_{ij,exp}^{RT}$, Γ_{exp}^{RT} і $k_{n,ij,exp}^{RT} = k_{ij,exp}^{RT}/k_{ij}^{RT}$, $\Gamma_{n,exp}^{RT} = \Gamma_{exp}^{RT}/\Gamma^{RT}$ зі змінами частот слів F і f, а також порівняння цієї поведінки з передбаченнями теорії. Раніше з цією метою було вивчено лише поведінку $k_{21,exp}^{RT}(F,f)$ і $k_{n,21,exp}^{RT}(F,f)$ [4]. Надалі індекс «*exp*», який наголошує на тому, що параметри здобуто шляхом симуляцій, а не підстановкою формул (2), ми будемо опускати. Ми генерували по 1000 текстів RT, у кожному з яких наявність або відсутність слова на деякій позиції позначали відповідно 1 і 0, а також було дотримано задані F, f і L. Далі було розраховано всі згадані вище ненормовані та нормовані параметри релевантності та виконано усереднення цих параметрів по 1000 текстах з однаковими F, f і L.

Як приклад, на Рис. 1 показано результати симуляцій для параметра Γ. Наші дані приводять до таких висновків. (**А**) Ненормовані параметри $\langle\Gamma^{RT}\rangle$ та ін. неявно залежать від частоти f (Рис. 1а), а відмінності кривих $k_{ij,exp}^{RT}(F)$ для різних f зростають для більших i, j. Це явище зникає лише в границі $f \to 0$ (див. формули (2)). Нормування параметрів на RT нівелює залежність від f (Рис. 1б), а тому вона повністю описується виразами (2). (**Б**) Ненормовані параметри не дорівнюють строго 1, 2, ≈ 0.27 тощо (див. вище), а лише прямують до цих значень при $F \to \infty$. Саму залежність від F вище не описано аналітично, проте її можна збагнути, враховуючи, що геометричний розподіл є лише наближенням строгішого дискретного розподілу ймовірності $p_{L,F}(\tau)$ [5]. (**В**) За означеннями, всі нормовані параметри, в т.ч.

$< \Gamma_n^{RT} >$, прямують до одиниці при $F \to \infty$ (Рис. 1б). Проте ця очікувана границя швидше досягається для параметрів, які є комбінаціями кумулянтів нижчих порядків. Інакше, наслідки скінченних розмірів L тексту сильніше виявляються для кумулянтів вищих порядків (див. також [2]). (**Г**) За центральною граничною теоремою, стандартне відхилення всіх параметрів спадає за степеневим законом $\sim F^{-a}$ ($a = 0.5$), хоча цей асимптотичний режим досягається при нижчих F для параметрів, що містять кумулянти нижчих порядків. Зокрема, для залежності $\Delta\Gamma^{RT}(F)$ одержуємо $a = 0.498$ (див. Рис. 1г).

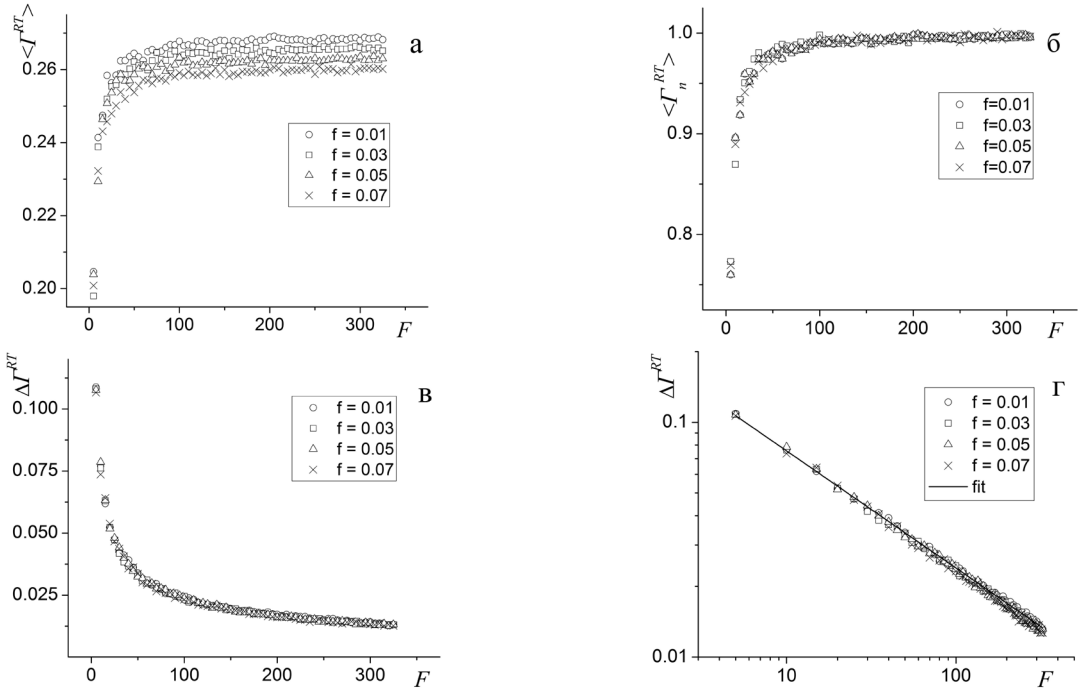

Рис. 1. (а) Залежності від частоти F параметрів релевантності $<\Gamma^{RT}>$ (а), $< \Gamma_n^{RT} >$ (б) і стандартного відхилення $\Delta\Gamma^{RT}$ у звичайному (в) і подвійному логарифмічному (г) масштабах. Дані усереднювали по 1000 текстах RT. Легенди описують дані для різних відносних частот *f*.

Для калібрування залежності емпіричного параметра $< R_n^{RT} > (F)$ автори [4] пропонували апроксимувати її та далі вносити поправки до параметра релевантності з урахуванням даних для RT. Зауважмо, що залежність $< R_n^{RT} > (F)$, подібну до $< \Gamma_n^{RT} > (F)$ на Рис. 1б, можна відтворити формулою $R_n^{RT} = \sqrt{F(L-F)/[(L+1)(F+2)]}$ [5], одержаною на підставі дискретного розподілу ймовірності $p_{L,F}(\tau) = \binom{L-\tau}{F-1} \Big/ \binom{L}{F}$. Він уточнює геометричний розподіл $p_{geo}(\tau)$ для скінченних L і F. Схожі вирази можна одержати та проаналізувати і для інших параметрів $k_{n,ij}^{RT}$, а також для Γ_n^{RT}. Такий аналіз буде предметом нашої наступної роботи.

Бібліографія

1. M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, A. M. Somoza, Europhys. Lett., **57**, 759 (2002).
2. J. P. Herrera, P. A. Pury, Eur. Phys. J. B, **63**, 135 (2008).
3. H. Zhou, G. W. Slater, Physica A, **329**, 309 (2003).
4. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, J. L. Oliver, Phys. Rev. E, **79**, 035102(R) (2009).
5. P. Carpena, P. A. Bernaola-Galván, C. Carretero-Campos, A. V. Coronado, Phys. Rev. E, **94**, 052302 (2016).