УДК 004.8

О.О. Арсірій, канд. техн. наук, доц.,
А.А. Чугунов, канд. екон. наук, доц.,
Ю.М. Ларченко, магістр,
Одес. нац. політехн. ун-т

ПОБУДОВА КОНТЕКСТНОЇ КАРТИ НА ОСНОВІ SOM ДЛЯ ВИДІЛЕННЯ КЛЮЧОВИХ СЛІВ ВЕБ-ДОКУМЕНТІВ ОСВІТНІХ ІНТЕРНЕТ-РЕСУРСІВ

О.О. Арсірій, А.А. Чугунов, Ю.М. Ларченко. **Побудова контекстної карти на основі SOM для виділення ключових слів веб-документів освітніх інтернет-ресурсів.** В результаті аналізу засобів просування освітніх інтернет-ресурсів запропоновано методику виділення ключових слів веб-документів на основі статистичних та морфологічних ознак шляхом побудови контекстної карти за допомогою SOM.

Ключові слова: веб-документ, ключові слова, СЯС, аналіз тексту, TF-IDF, самоорганізовані карти Кохонена, контекстні карти

Е.А. Арсирий, А.А. Чугунов, Ю.Н. Ларченко. **Построение контекстной карты на основе SOM для выделения ключевых слов веб-документов образовательных интернет-ресурсов.** В результате анализа средств продвижения образовательных интернет-ресурсов предложена методика выделения ключевых слов веб-документов на основе статистических и морфологических признаков путем построения контекстной карты с помощью SOM.

Ключевые слова: веб-документ, ключевые слова, СЯС, анализ текста, TF-IDF, самоорганизующиеся карты Кохонена, контекстные карты.

E.A. Arsiriy, A.A. Chugunov, Y.N. Larchenko. **Creation contextual map based on SOM to searched keywords of web documents of educational internet-resources.** As a result of the analysis of tools of advancement internet resources offered the technique of searched keywords of web documents based on statistical and morphological characters by creating the contextual map based on SOM.

Keywords: web-document, the keywords, SCS, text analysis, TF-IDF, Kohonen self-organizing maps, contextual map.

Актуальність проблеми. Постановка задачі. Специфіка об'єкта управління, слабкий розвиток інформаційних систем та загострення проблеми профорієнтації в освітньої галузі роблять актуальним завдання розробки відвідуваних інтернет-ресурсів кафедр вищих навчальних закладів [1]. Розробка освітнього інтернет-ресурсу (ОІР) складається із етапів побудови, аналізу і просування в мережі Інтернет. Інформаційно-представницька спрямованість ОІР посилює важливість етапу просування, професійне виконання якого забезпечує оперативний доступ до довідкової, практичної і теоретичної інформації викладачів, студентів, абітурієнтів та інших зацікавлених осіб, підвищує кількість і якість відвідувачів сайту. Більшість потенційних відвідувачів потрапляють на ОІР при переході за посиланнями у результатах пошукових систем (ПС), робота яких заснована на визначенні так званих «ключових слів» (КС) веб-документу. З переліку таких слів складається семантичне ядро сайту (СЯС). СЯС являє собою список КС і їх комбінацій, записаних в метатегах *keywords* і розподілених в контенті сайту, а саме, у тезі *title*, в *alt*-аттрибутах, в тексті внутрішніх і зовнішніх посилань, у виділеннях жирним і курсивним шрифтом, на початку контенту сайту, в назві файлів, в *URL* та ін. Від повноти і точності розробки СЯС залежить положення сайту в списку видач ПС [2]. Розробка СЯС є основою так званих «білих» методів пошукової оптимізації, які, в свою чергу, є єдино можливими в умовах обмеженості грошового бюджету. Наявність обмежуючих умов ускладнює етап просування ОІР на відміну від просування конкуруючих з ними комерційних інформаційно-представницьких інтернет-ресурсів.

ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

З іншого боку тривалість розробки СЯС, а саме виділення КС веб-документів збільшується за рахунок наявності динамічного контенту в ОІР, а саме ведення тематичних форумів, та новинних блогів, періодичною зміною назв й переліків спеціальностей, появою нових наукових праць й видань і таке інше При цьому тривалість виділення КС веб-документів може значно затримувати необхідну періодичність оновлення СЯС, що призводить до зниження повноти і точності СЯС, і ресурс втрачає свої позиції на сторінці видачі пошукових результатів. Тому метою дослідження є побудова методики автоматизованого виділення КС у веб-документі для скорочення часу розробки СЯС ОІР з динамічним контентом без втрати повноти і точності. Для розробки методики необхідно: проаналізувати існуючі методи аналізу тексту для виділення ознак КС веб-документів; визначити спосіб представлення виділених ознак наборів слів для автоматизації їх обробки; проаналізувати можливості SOM як засобу інтелектуальної візуалізації матриць даних при виділенні КС; запропонувати методику реалізації виділення КС за допомогою SOM на основі даних аналізу тексту.

Методи аналізу тексту для виділення КС. Всю сукупність представлених на сьогоднішній день методів аналізу тексту, відносно задачі виділення КС, можна умовно розділити на дві групи: лінгвістичний та статистичний аналіз.

Використання методів лінгвістичного аналізу дає можливість отримати сенс тексту за його семантичною структурою, що дозволяє точніше аналізувати текст, виділяючи його структурні особливості, але є більш трудомістким і складним у використанні, крім того передбачає орієнтацію на конкретну мову з її конкретними семантичними особливостями, що обумовлює погану міжмовну переносимість. Складність використання методів лінгвістичного аналізу для автоматизації виділення КС веб-документів пов'язана з багатством семантики та морфології природних мов. При цьому формальний опис правил природної мови та їх реалізація потребує залучення фахівців з області лінгвістики. Реалізація методів лінгвістичного аналізу у виділенні КС можлива за рахунок врахування морфологічних особливостей слова як елемента тексту (довжина слів, частини мови, рід, власна чи загальна назва, відміна, форма порівняння тощо).

Використання методів статистичного аналізу дає можливість виділити КС по частотному розподілу слів у тексті не залежно від мови, але не дозволяє повною мірою врахувати сенс тексту. При цьому автоматизація виділення КС веб-документів на основі використання статистичних методів більш проста та зручна у реалізації.

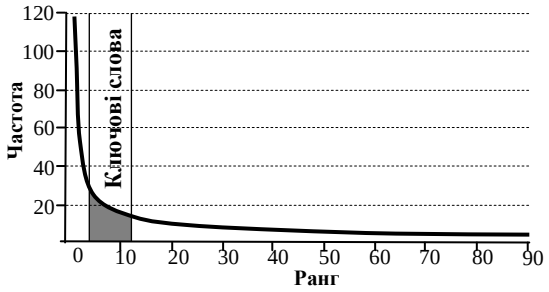
Методи статистичного аналізу для виділення КС базуються на емпіричних законах Ципфа [3]. Відповідно до першого закону Ципфа можна отримати графік залежності рангу r від частоти n (рис. 1, а). Кількість входження слова t у документ d буде його частотою n_t. Якщо розташувати частоти слів по мірі їх убування і пронумерувати, тоді порядковий номер частоти буде рангом слова t — r_t. Дослідження показують, що найбільш значимі для документу слова лежать у середній частині графіка. Від вибору значень порогу зверху та знизу для обчислення рангу значимих слів залежить точність, якість і кількість виділення КС. Якщо встановити високе значення порогу зверху та низьке знизу, то в КС потраплять допоміжні слова, якщо навпаки — то можлива втрата смислових термінів. Тому використання тільки такого підходу для виділення КС викликає практичні складнощі.

Використання для визначення важливості слова веб-документу, який є частиною колекції веб-документів ОІР статистичної міри *TF-IDF* (від англ. *TF — term frequency, IDF — inverse document frequency*) — є продовженням робіт Ципфа і ліквідує складнощі їх практичного використання. При розрахунку *TF-IDF* вага слова *t* пропорційна кількості вживання цього слова в документі *d*, і обернено пропорційна частоті вживання слова в інших документах колекції *D*.

Таким чином, для визначення слів з високою вагою, які зустрічаються у досліджуваному документі необхідно визначити *TF* як відношення числа разів появи слова *t* до довжини документа *d*

$$TF(t,d) = \frac{n_t}{N}, \qquad (1)$$

де n_t — частота слова t у документі d;
N — кількість слів у документі d.



Слова	Показов в месяц
экономическая кибернетика	1117
кафедра экономической кибернетики	103
методы экономической кибернетики	77
экономическая кибернетика специальность	69
экономическая кибернетика скачать	54
экономическая кибернетика учебник	30
экономическая кибернетика работа	25
факультет экономической кибернетики	21
экономическая кибернетика лекции	21
экономическая кибернетика реферат	20
основные понятия экономической кибернетики	14
экономическая кибернетика вакансии	12

а *б*

Рис. 1. Можливості виділення КС: за першим законом Ципфа (а), згідно сервісу Wordstat.yandex (б)

Для пониження ваги поширених слів, які зустрічаються майже у всіх документах колекції, вводять інверсну частоту IDF як відношення числа всіх документів колекції D до числа документів що містять слово t. Необхідно зазначити, що IDF буде близькою до нуля для поширених слів

$$IDF(t,D)=\frac{D}{d}, \tag{2}$$

де D — загальна кількість документів колекції;
d — кількість документів, в яких зустрічається слово t.
Статистична міра TF-IDF є добутком двох співмножників:

$$TFIDF(t,d,D)=TF(t,d)\times IDF(d,D). \tag{3}$$

Сучасні засоби статистики ПС (*wordstat.yandex, adstat.rambler, adwords.google*) за допомогою власних алгоритмів також дозволяють отримати список КС асоційованих з веб-документом на основі пошукових запитів користувачів мережі Інтернет (рис. 1, б).

Представлення статистичних і морфологічних ознак наборів слів для автоматизації їх обробки. Для автоматизації обробки веб-документу за статистичним та морфологічним аналізом набір слів можна представити у вигляді числових бінарних характеристичних векторів $\{\vec{X}_1,\vec{X}_2,...,\vec{X}_n\}$, де $\vec{X}_t$ — характеристичний вектор слова t, а n — кількість слів в обраному для представлення фрагменті тексту. В свою чергу бінарний характеристичний вектор слова $\vec{X}_t$ складається з статистичних і морфологічних ознак слова $\vec{X}_t=\{x_1,x_2,...,x_k,y_1,y_2,...,y_m\}$, $x=\{0,1\}$, $y=\{0,1\}$, де x, y — оцінки відповідно статистичної і морфологічної характеристики слова; k, m — кількість статистичних і морфологічних характеристик відповідно.

Значення оцінки характеристики може дорівнювати одиниці, якщо така властивість притаманна слову, або нулю, якщо слово такої властивості не має. Кількість характеристик визначається експертом. Набір характеристичних векторів можливо представити в вигляді

ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

матриці даних в випадку якщо для всіх слів відомі значення всіх ознак, які їх характеризують. Кожен зі стовпців матриці відповідає одному слову із набору (табл. 1).

SOM як засіб інтелектуальної візуалізації матриць даних при виділенні КС. Для реалізації процедур інтелектуальної візуалізації матриць даних пропонується використовувати нейромережевий підхід, заснований на застосуванні самоорганізованих карт Кохонена (self-organizing map – SOM) [4]. Виділяють два методи візуалізації відображення вхідного простору в простір ознак за допомогою побудови самоорганізованих карт:

— у вигляді гнучкої решітки взаємозв'язаних нейронів, де вектори вагових коефіцієнтів є покажчиками, які спрямовані від відповідних нейронів карти до об'єктів із вхідного простору;

— у вигляді когерентних областей, які утворюють однаково помічені нейрони. При цьому мітки класів призначаються нейронам карти в залежності від їх збудження представниками вхідного простору, а під когерентністю розуміється те, що кожна із груп нейронів є відособленою множиною безперервних міток.

SOM які реалізовані за другим методом візуалізації називають контекстними або семантичними. Процес побудови контекстної SOM складається із процедур самоорганізації і градуювання. При виконанні процедури самоорганізації вектори X обираються із табл. 1 випадково і діють як набір значень на вході SOM (рис. 2). Вибірка і наступна за нею адаптація SOM продовжується ітераційно до поки отриманий асимптотичний стан SOM не буде визнаний стаціонарним. При виконанні процедури *градуювання* нейронам SOM присвоювалися мітки згідно найкращої відповідності векторів вагових коефіцієнтів нейронів вхідним векторам *X*.

Таблиця 1

Матриця вхідних даних

Ознака	Слова					
	X_1	X_2	...	X_t	...	X_n
x_1	1	0	...	0	…	1
x_2	0	0	…	1	…	0
...	…	…	…	…	…	…
x_k	0	1	…	0	...	0
y_1	0	0	...	1	…	1
y_2	1	0	…	1	…	0
...	…	…	…	…	…	…
y_m	1	0	…	0	...	1

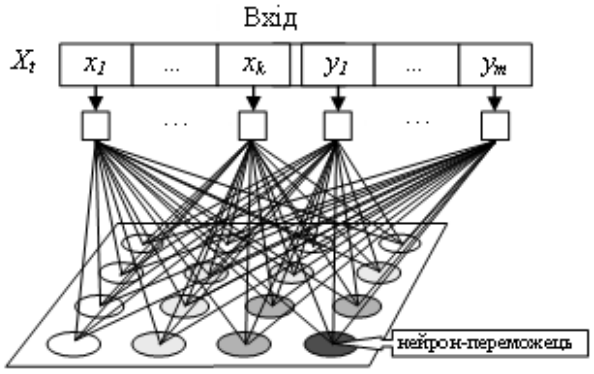


Рис. 2. Відображення вхідного вектора X_t на SOM

Методика автоматизованого виділення КС веб-документу. Для виділення КС у веб-документах пропонується використати комплексний аналіз, який базується на поєднанні можливостей статистичного і лінгвістичного аналізу і використовує контекстну SOM як засіб інтелектуальної візуалізації. Запропонована методика складається з наступних етапів.

Етап 1. Розрахунок міри TF×IDF та отримання статистики ПС.

Етап 2. Отримання морфологічних характеристик елементів тексту.

Етап 3. Представлення статистичних і морфологічних характеристик елементів фрагменту веб-документу у вигляді матриці вхідних даних (див. табл. 1).

Етап 4. Побудова контекстної SOM на основі матриці вхідних даних для отримання КС.

Покажемо реалізацію запропонованої методики на прикладі фрагменту тексту, що знаходиться на головній сторінці сайту кафедри ЕКІТ ОНПУ (http://ecit.od.ua/index.php).

Етап 1. Розрахунок міри TF×IDF для елементів текстового фрагменту виконано за формулами (1)…(3). Для розрахунку використано наступні константи: довжина документу *N* дорівнює 677 слів; загальна кількість документів *D* у базі дорівнює 8 млрд. документів. При цьому за базу прийнято кількість веб-документів проіндексованих ПС Google у російському сегменті мережі Інтернет. За допомогою сервісів ПС (див. рис. 1, б) були отримані рекомендації

щодо статистики КС. На основі отриманих статистичних ознак для подальшого аналізу з метою виділення КС із всіх елементів текстового фрагменту було обрано 16, які показано в табл. 2.

Етап 2. Залученим експертом для статистично обраних елементів запропоновані морфологічні ознаки у вигляді значення довжини слова і назви частини мови (табл. 2).

Таблиця 2

Перелік статистичних і морфологічних ознак слів фрагменту тексту

	Слово (*t*)	Статистичні ознаки						Морфологічні ознаки	
		n	*r*	*d*	TF	IDF	TF×IDF	Довжина	Частина мови
1	Кафедра	16	3	0,0147	0,024	544,218	12,862	16	іменник
2	Экономическая	29	2	0,0277	0,043	288,809	12,371	29	прикметник
3	Кибернетика	13	5	0,00304	0,019	2631,579	50,533	13	іменник
4	И	30	1	5,11	0,044	1,566	0,069	30	сполучник
5	Информационный	16	3	0,0518	0,024	154,440	3,650	16	прикметник
6	Технологии	10	6	0,235	0,015	34,043	0,503	10	іменник
7	Выполняет	2	10	0,248	0,003	32,258	0,095	2	дієслово
8	Подготовка	5	7	0,0818	0,007	97,800	0,722	5	іменник
9	Бакалавр	2	10	0,00495	0,003	1616,162	4,774	2	іменник
10	По	14	4	3,46	0,021	2,312	0,048	14	прийменник
11	Направление	3	9	0,0606	0,004	132,013	0,585	3	іменник
12	Область	3	9	0,364	0,004	21,978	0,097	3	іменник
13	Знания	2	10	0,0625	0,003	128,000	0,378	2	іменник
14	Экономика	4	8	0,155	0,006	51,613	0,305	4	іменник
15	Предпринимательство	1	11	0,0932	0,001	85,837	0,127	1	іменник
16	Специалист	3	9	0,0718	0,004	111,421	0,494	3	іменник
17	Специальность	3	9	0,0182	0,004	439,560	1,948	3	іменник
18	Магистр	2	10	0,0852	0,003	93,897	0,277	2	іменник

Етап 3. Для подальшої реалізації процедур інтелектуальної візуалізації за допомогою SOM данні із табл. 2 необхідно представити у вигляді числових бінарних векторів (див. табл. 1). При цьому в якості вектору $\vec{X}_t$ виступає слово *t* із обраних 16 слів. Числові і символьні значення статистичних і морфологічних ознак трансформуються в категоріальні значення, яким присвоюється двійкове число. Довжина двійкового числа дорівнює кількості обраних категорій для кодування відповідної ознаки слова *t*.

Етап 4. Для побудови контекстної SOM з метою отримання КС у NNTools ППП Matlab було створено нейронну мережу *net=newsom*([0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;0 1;],[10 10], *'gridtop'),* розміром 10×10 нейронів і на протязі 20000 епох навчання *net.trainParam.epochs=20000* здійснено процедуру самоорганізації за допомогою функції *net=train(net,P),* за вхідний параметр *P* взято матрицю вхідних даних представлену у вигляді двійкових чисел. Градуювання SOM виконано за допомогою послідовного виконання функції *plotsomhits (net,test),* для всіх векторів стовпців *test* матриці вхідних даних. На контекстній SOM, яку сформовано після градуювання, ясно видно геометричний порядок у вигляді когерентних областей, отриманий для різних категорій КС (рис. 3).

Висновки. Використання запропонованої методики автоматизованого виявлення КС у веб-документі на основі статистичних та морфологічних ознак шляхом побудови контекстної SOM дозволило без втрати повноти і точності скоротити на 25 % час оновлення СЯС з динамічним контентом кафедри ЕКІТ ОНПУ, що дає змогу утримувати високу позицію сайту на сторінці видачі результатів ПС і збільшити кількість відвідувачів сайту — потенційних студентів кафедри. Запропонована методика є універсальною, використання її особливо актуально для просування в умовах обмеженості грошового бюджету подібних розглянутому ОІР, інших веб-ресурсів.

ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

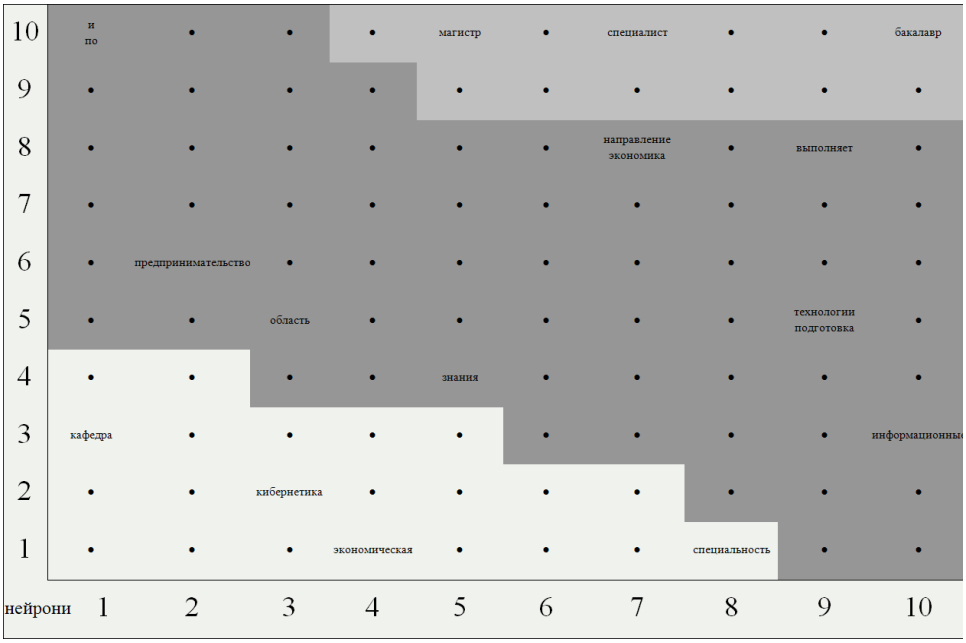


Рис. 3. Контекстна SOM із трьох когерентних областей

Література

1. Солдатов, А.В. Информационная система как основа эффективного управления вузом / А.В. Солдатов // Университетское управление: практика и анализ. — 2004. — № 2(31). — С. 116 — 119.
2. Арсирий, Е.А. Автоматизация разработки и обновления семантического ядра сайта с динамическим контентом / Е.А. Арсирий, С.Г. Антощук, О.А. Игнатенко, Б.Ф. Трофимов // Искусственный интеллект. —2012. — № 4. — С. 464 — 473.
3. Закон Ципфа — Вводная статья [Электронный ресурс] / Режим доступа: http://webpavilion.ru/статьи/закон-ципфа-вводная. — 21.12.2012.
4. Кохонен, Т. Самоорганизующиеся карты / Т. Кохонен: Пер. 3-го англ. изд. — М.: БИНОМ, Лаборатория знаний, 2008. — 655 с.

References

1. Soldatov, A.V. Informacionnaja sistema kak osnova jeffektivnogo upravlenija vuzom [Information system as the basis for effective university management]/ A.V.Soldatov // Universitetskoe upravlenie: praktika i analiz [University Management: Practice and Analysis]. — 2004. — # 2(31). — pp. 116 — 119.
2. Arsiriy, E.A. Automation of development and updating for semantic kernel of a site with dynamic content / E.A. Arsiriy, S.G. Antotshuk, O.A. Ignatenko, B.F. Trofimov // Artificial intelligence. — 2012. — # 4. — pp. 464 — 473.
3. Zakon Cipfa — Vvodnaja stat'ja [Zipf's Law — An introductory article] [Electronic resource] / Available at: http://webpavilion.ru/stat'i/zakon-cipfa-vvodnaja. — 21.12.2012.
4. Kohonen, T. Samoorganizujushhiesja karty [Self-organizing map] / T. Kohonen: Per. 3-go angl.izd [Transl. 3-rd edition from Engl.] . — Moscow, 2008. — 655 p.

Рецензент д-р техн. наук, проф. Одес. нац. політехн. ун-ту Крисилов В.А.

Надійшла до редакції 21 грудня 2012 р.