УДК 519.765:519.767:004.89

Б.М. Павлишенко

Львівський національний університет ім. Івана Франка, Львів

ЧАСОВА ЗАЛЕЖНІСТЬ КВАНТИТАТИВНИХ ХАРАКТЕРИСТИК КЛЮЧОВИХ ТЕГІВ У RSS КАНАЛАХ

Побудована теоретична модель квантитативних семантичних характеристик ключових тегів у RSS каналах. На прикладі тестових масивів RSS каналів експериментально показано, що динаміка квантитативних ознак ключових тегів відображає тенденції у семантично заданій тематиці RSS каналів.

Ключові слова: *ключові теги, RSS канали, семантичні поля, інтелектуальний аналіз текстів.*

Вступ

Використання методик інтелектуального аналізу даних текстового типу є ефективним у багатьох напрямках сучасних інформаційних технологій [1, 2]. Більшість сучасних інформаційних сайтів використовують RSS формат для виведення коротких інформаційних повідомлень. Такий формат є зручний для аналізу, оскільки він містить інформаційні повідомлення у структурованому вигляді із виділенням джерел, авторів, посилань, заголовків та основного тексту короткого повідомлення. Використання такого формату є зручним у спеціалізованих програмах для читання RSS-стрічок. Такі програми дають можливість переглядати одночасно короткі повідомлення про новини із різних сайтів. RSS канали є також зручним об'єктом інтелектуального аналізу, оскільки короткі повідомлення є семантично насиченими відповідними ключовими тегами. Під ключовими тегами будемо розглядати лексеми, які найбільш точно характеризують суть інформаційного повідомлення. Цікавим, на наш погляд, є аналіз прогностичної ефективності часових залежностей ключових тегів у повідомленнях інформаційних агенств.

Аналіз останніх досліджень. У роботі [3] проводився кластерний аналіз текстових масивів у просторі семантичних полів. На основі теорії алгебраїчних решіток проаналізовано семантичні зв'язки між ключовими поняттями у текстових масивах. Запропонована модель кластеризації текстових документів у семантичному просторі дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки. Такий структурний поділ відображає класифікацію документів за новими ознаками документів, зокрема за авторством текстів. У роботі [4] розглянута теоретико-множинна концепція семантичних полів у масивах текстових даних. Семантичні класи утворюються як відношення еквівалентності. Семантичне поле визначається як сегмент, який утворюється семантичними класами, об'єднаними бінарним кластером у структурному відношенні семантичного розбиття лексемного словника текстових масивів. Розглянуто відношення, яке описує розбиття словника на семантичні класи зі структурою, яка визначає семантичні поля лексемного словника. Проаналізовано утворення семантичних полів на основі лексемних відношень, зокрема, таких як сполучення у тексті лексем семантичного поля та лексем полеутворюючої множини. Використання концепції семантичних полів є ефективним у векторній моделі текстових документів внаслідок зменшення розмірності фазового простору представлення документів. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстових даних. У роботі [5] розглядається сингулярний розклад матриці семантичних ознак типу "частоти_семантичних_полів документи", який дає можливість аналізувати текстові документи у новому просторі семантичних концептів. Ієрархічна кластеризація документів у такому просторі відображає класифікаційну структуру документів за різними ознаками, зокрема за авторством текстів. Розмірність простору семантичних концептів визначається рангом апроксимації матриці семантичних ознак при сингулярному розкладі і може бути суттєво меншою за розмірність простору семантичних полів.

Постановка задачі. Побудуємо теоретико-множинну модель ключових тегів заданого семантичного поля у масиві RSS каналів. Дослідимо кількісні характеристики появи ключових тегів заданого семантичного поля у RSS каналах світових інформаційних агентств. На основі виявлених характеристик проаналізуємо їх часову динаміку з точки зору аналізу наявних тенденцій у семантично заданій тематиці інформаційного простору.

Теоретична модель

Розглянемо множину RSS стрічок, які завантажуються із сайтів у такому вигляді

$$RSS = \left\{ \; rss_{lt} \; | \, l = 1,2,...N_l^{rss}, t = 1,2,...N_t^{rss} \right\} , \quad (1)$$

де rss_{lt} – інформаційне повідомлення з l-го сайту, завантажене у часовий момент t; N_l^{rss} – кількість інформаційних сайтів, з яких завантажуються RSS-стрічки; N_t^{rss} – кількість завантажень у задані часові моменти. Оскільки сайти вибирають за заданою тематикою, то їх RSS-канали можна розглядати як єдиний тематичний масив об'єднаного семантичного спектру. Сформуємо деяке семантичне поле лексем, які будуть відображати задану тематику аналізу. Семантичні поля формуються на основі експертного аналізу, одні і ті ж лексеми можуть одночасно належати до різних семантичних полів. Під семантичним полем розуміють таку множину лексем, які об'єднані деяким спільним поняттям [6, 7]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та інші.

Введемо відображення лексемного складу словника W на множину семантичних полів S за допомогою деякого оператора U_{ws}

$$U_{ws} : w_i \to s_k, \; i = 1,2..., N_w; k = 1,2..., N_s . \quad (2)$$

Оператор U_{ws} задамо таблицею, яка визначається експертним лексикографічним аналізом. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2..., N_w \right\}. \quad (3)$$

Розглянемо кількісні характеристики тегів у масивах RSS-каналів. Однією з найбільш поширених текстових характеристик є частота лексеми у текстах, яку розглядають як відношення кількості входжень цієї лексеми у заданий текстовий об'єм до загальної кількості лексем у цьому текстовому фрагменті. Оскільки RSS-канали є семантично насиченими текстовими об'єктами малих розмірів, як частотну характеристику лексеми будемо розглядати кількість RSS-стрічок у заданому списку, які містять цю лексему. Ведемо мультимножину ключових тегів, які зустрічаються у RSS-каналах

$$KW = \left\{ n_{jt}^{rss}(kw_j) \mid j = 1,2..., N_{kw}^{rss}, t = 1,2,...N_t^{rss} \right\}, \quad (4)$$

де n_{jt}^{rss} – кількість входжень ключового тега kw_j в лексемному складі множини RSS-стрічок rss_{lt} завантажених одночасно в деякий момент часу t. Величину n_{jt}^{rss} можна розглядати як

$$n_{jt}^{rss} = \sum_{l=1}^{N_l^{rss}} f_{rss}(kw_j, rss_{lt}), \quad (5)$$

де

$$f_{rss}(kw_j, rss_{lt}) = \begin{cases} 1, kw_j \in rss_{lt}; \\ 0, kw_j \notin rss_{lt}. \end{cases}$$

Різні ключові теги можуть мати різну кількість входжень у заданому масиві RSS-стрічок, тому виникає необхідність у введенні такої кількісної характеристики, яка б давала можливість порівнювати часову динаміку різних лексем. Однією з таких характеристик може бути відносна частота лексеми, яка визначається відношенням частоти лексеми до частоти цієї лексеми усередненої по всіх часових проміжках:

$$(n_{jt}^{rss})_{relative} = \frac{n_{jt}^{rss}}{(n_{jt}^{rss})_{avg}}, \quad (6)$$

де

$$(n_{jt}^{rss})_{avg} = \frac{\sum_{t=1}^{N_t^{rss}} n_{jt}^{rss}}{N_t^{rss}} .$$

Для визначення та аналізу інформаційних тенденцій у новинах необхідно ввести диференційні характеристики, які б відображали часові зміни. Однією із таких характеристик може бути різниця між відносними частотами ключового тега у двох масивах RSS-стрічок завантажених у два послідовних проміжки часу:

$$(dn_{jt}^{rss})_{relative} = (n_{jt}^{rss})_{relative} - (n_{jt-1}^{rss})_{relative} . \quad (7)$$

Поряд з аналізом динаміки частот лексем за часовою послідовністю завантажених масивів RSS-стрічок, актуальним є аналіз, проведений для згрупованих та усереднених даних за деяким часовим проміжком, наприклад, по днях. Відносні частоти згрупованих та усереднених даних можна розглядати як

$$(n_{jm}^{rss})_{relative}^{group} = \frac{(n_{jm}^{rss})_{avg}^{group}}{(n_{jt}^{rss})_{avg}}, \; m = 1,2,...N_g^{rss}, \quad (8)$$

де

$$(n_{jm}^{rss})_{avg}^{group} = \frac{\sum_{t=1}^{N_{tm}^{rss}} n_{jt}^{rss}}{N_{tm}^{rss}}, \quad (9)$$

де N_g^{rss} – визначає кількість груп, при групуванні по днях ця величина рівна кількості аналізованих днів; N_{tm}^{rss} – кількість завантажень заданого списку RSS-стрічок у групі m, тобто в день m. У різні дні може бути різна кількість завантажень. Різницю відносних частот згрупованих даних можна розрахувати так

$$(dn_{jt}^{rss})_{relative}^{group} = (n_{jm}^{rss})_{relative}^{group} - (n_{jm-1}^{rss})_{relative}^{group} . \quad (10)$$

Експериментальна частина

Для апробації запропонованої теоретичної моделі був розроблений пакет прикладних програм на мові С.

Для завантаження масивів RSS-стрічок використовувались бібліотеки curl, libxml. Для обробки RSS-стрічок використовувалися функції бібліотеки libxml:

```
xmlNodePtr xmlDocGetRootElement(xmlDocPtr doc);
xmlChar * xmlNodeGetContent (xmlNodePtr cur).
```

Виділені із RSS-стрічок ключові теги зберігалися та оброблялися за допомогою функцій бібліотеки бази даних sqlite3:

```
int sqlite3_open( const char *filename, sqlite3 **ppDb);
int sqlite3_prepare_v2( sqlite3 *db, const char *zSql,
int nByte, sqlite3_stmt **ppStmt, const char **pzTail); int
sqlite3_step(sqlite3_stmt*);
int sqlite3_exec( sqlite3*, const char *sql, int
(*callback)(void*,int,char**,char**), void *, char
**errmsg).
```

Масив RSS-стрічок сформований на основі RSS-каналів таких сайтів: cnn.com, yahoo.com, reuters.com, newyorkfed.org, ft.com та інших, всього 132 канали. Для аналізу ключових тегів було вибрано семантичне поле загальної економічної тематики, у яке ввійшли такі теги – dollar, economic (economics, economical), investors, market, money, trading, sales, business та інші, всього 30 ключових тегів. На рис. 1 наведена часова залежність кількості появ тегів у RSS-каналах. На рис. 2 наведена часова залежність відносної кількості появ тегів у RSS-каналах. На рис. 3 наведені згруповані по днях відносні кількості появ тегів. На рис. 4 наведені часові залежності змін відносної кількості появ тегів. На рис. 5 наведені згруповані по днях зміни відносних кількостей появ тегів.

Рис. 1. Часова залежність кількості появ у RSS-каналах тегів market, business, economic у порядку спадання частоти. По осі X відкладено порядкові номери завантажень у порядку спадання часу завантажень, а по осі Y – кількості появ тегів

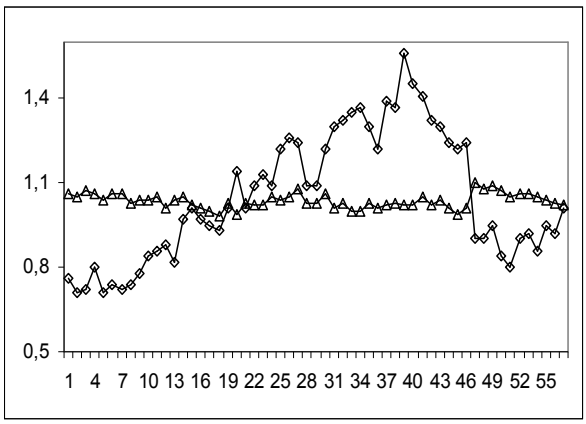

Рис. 2. Часова залежність відносної кількості появ у RSS-каналах тегів market (Δ) та economic (◊) . По осі X відкладено порядкові номери завантажень в порядку спадання часу завантажень, а по осі Y – відносні кількості появ тегів

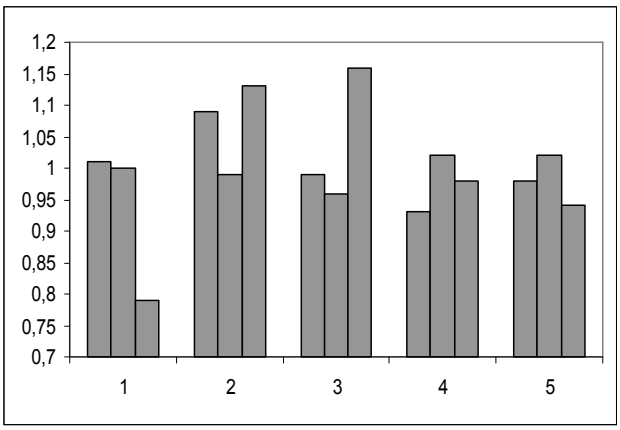

Рис. 3. Згруповані по днях відносні кількості появ тегів business, market, economic (стовпці зліва направо на гістограмі). По осі X відкладено дні у порядку спадання, а по осі Y – відносні кількості появ тегів усереднені по днях

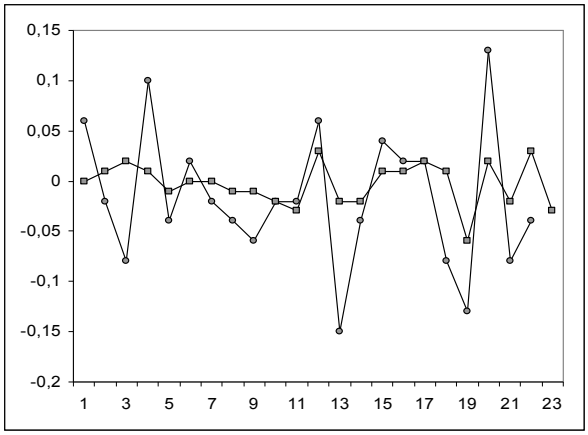

Рис. 4. Часова залежність зміни відносної кількості появ у RSS-каналах тегів business (□) та economic (○). По осі X відкладено порядкові номери завантажень у порядку спадання часу завантажень, а по осі Y – відносної кількості появ тегів

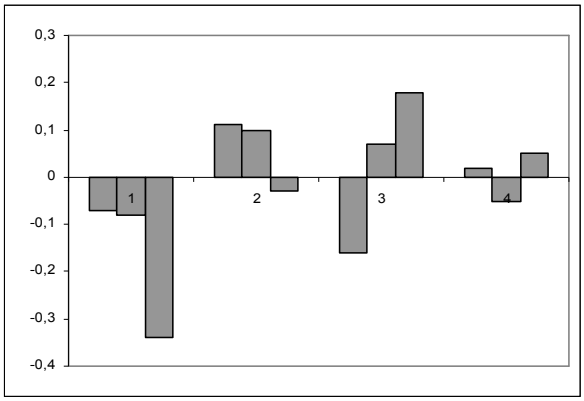

Рис. 5. Згруповані по днях зміни відносних кількостей появ тегів trading, business, economic (стовпці зліва направо на гістограмі). По осі X відкладено дні у порядку спадання, а по осі Y – відносні кількості появ тегів усереднені по днях

На основі аналізу динаміки кількісних характеристик ключових тегів можна побачити, що найбільш чутливою до змін є часова залежність різниці частот ключових тегів розрахованих для двох послідовних часових проміжків. У випадку аналізу відносних частот можна порівнювати на одному графіку зміни для різних ключових тегів. Зміни числа ключових тегів у періодично завантажених масивах RSS стрічок відображають миттєві та швидкі зміни інформаційних трендів. Зміни числа ключових тегів, які згруповані та усереднені по днях, відображають більш фундаментальні тенденції інформаційного простору.

Висновки

Запропонована теоретична модель, яка описує динаміку ключових тегів у масиві RSS-каналів. Ключові теги розглядаються як лексеми деякого семантичного поля заданої тематики. Як кількісні характеристики розглядаються відносні частоти лексем та різниці цих частот у послідовно завантажених масивах RSS-стрічок. Також розглядаються ці ж величини згруповані та усереднені по днях завантажень RSS-каналів. Розроблено відповідне програмне забезпечення та проведена експериментальна апробація запропонованої моделі. Отримані результати показують існування часової динаміки квантитативних характеристик ключових тегів. Різні теги мають різний характер та період коливань. Отримані результати дають можливість припустити, що у часовій динаміці параметрів тегів відображаються інформаційні тенденції новин, для яких ці теги є семантично значимими. Отже, вибираючи ключові теги із множини лексем семантичного поля заданої тематики, можна відслідковувати загальні тренди новин, тематика яких визначається заданим семантичним полем.

Список літератури

1. Анализ данных и процессов: учеб. пособие / А.А. Брасегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. – СПб.: БХВ-Петербург, 2009. – 512 с.:ил.

2. Pantel P. From Frequency to Meaning: Vector Space Models of Semantics / Patrick Pantel, Peter D. Turney // Journal of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141-188.

3. Павлишенко Б.М. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів / Б.М. Павлишенко // Електроніка та інформаційні технології. – 2011. – Вип. 1. – С. 212-222.

4. Павлишенко Б.М. Використання концепції семантичного поля у векторній моделі текстових документів / Б.М. Павлишенко // Східно-Європейський журнал передових технологій. – 2011. – № 6/2 (54). – С. 7-11.

5. Павлишенко Б.М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів/ Б.М. Павлишенко // Математичні машини і системи. – 2012. – № 1. – С. 69-76.

6. Вердиева З.Н. Семантические поля в современном английском языке / З.Н. Вердиева – М.: Высшая школа, 1986. – 120 с.

7. Левицкий В.В. Экспериментальные методы в семасиологии / В.В. Левицкий, И.А. Стернин. – Воронеж: Изд-во ВГУ, 1989. – 192 с.

Надійшла до редколегії 7.03.2012

Рецензент: д-р фіз.-мат. наук, проф. О.М. Бордун, Львівський національний університет імені Івана Франка, Львів.

ВРЕМЕННАЯ ЗАВИСИМОСТЬ КВАНТИТАТИВНЫХ ХАРАКТЕРИСТИК КЛЮЧЕВЫХ ТЕГОВ В RSS КАНАЛАХ

Б.М. Павлышенко

Построена теоретическая модель квантитативных семантических характеристик ключевых тегов в RSS каналах. На примере текстовых массивов RSS каналов экспериментально показано, что динамика квантитативных признаков ключевых тегов отображает тенденции в семантически заданной тематике RSS каналов.

Ключевые слова: *ключевые теги, RSS каналы, семантические поля, интеллектуальный анализ текстов.*

TIME DEPENDENCE OF QUANTITATIVE CHARACTERISTICS OF KEY TAGS IN RSS FEEDS

B.M. Pavlyshenko

The theoretical model of quantitative characteristics of key semantic tags in RSS feeds haS been suggested. On the example of test arrays of RSS feeds it has been shown experimentally that the dynamics of quantitative characteristics of key tags represents the trends in given semantic thematics of RSS feeds.

Keywords: *key tags, RSS feeds, semantic fields, text mining.*