

УДК 519.765:519.767:004.93

АНАЛІЗ МІКРОБЛОГІВ КОРИСТУВАЧІВ НА ОСНОВІ ГРАТКИ СЕМАНТИЧНИХ КОНЦЕПТІВ

Б.М. Павлишенко

Львівський національний університет імені Івана Франка

e-mail:pavlsh@yahoo.com

Інтелектуальний аналіз веб-контенту є складовою частиною сучасних інформаційних технологій. Система мікроблогів Twitter є одним із популярних засобів взаємодії користувачів за допомогою коротких повідомлень. Лаконічність Twitter-повідомлень зумовлює високу густину тематично значимих ключових слів і створює перспективність досліджень мікроблогів засобами інтелектуального аналізу. Інтелектуальний аналіз слабо структурованих даних є ефективним методом дослідження текстових масивів [1,2]. Одним із перспективних напрямків аналізу даних є теорія аналізу формальних концептів [2,3,4]. В цій теорії розглядається відношення об'єктів та їх атрибутів, на основі якого будують алгебраїчну ґратку формальних концептів. Кожен концепт об'єднує множину об'єктів та їх спільних атрибутів. На основі частих множин спільних атрибутів виявляють асоціативні правила, які відображають зв'язки між атрибутами на множині аналізованих об'єктів. Актуальним на даний час є створення моделі формальних концептів для аналізу мікроблогів, яка б враховувала семантичну структуру повідомлень. Для цього доцільно ввести поняття семантичного поля, яке б об'єднувало ключові лексеми тематики аналізу [5]. Розглянемо теоретико-множинну модель, яка описує повідомлення мікроблогів, згрупованих за користувачами. В дослідженнях використаємо аналіз формальних концептів, який базується на теорії алгебраїчних ґраток [2,3,4].

Розглянемо модель, яка описує повідомлення мікроблогів, їх словник, користувачів та тематичні групи. Загальний словник аналізованого масиву повідомлень розглянемо як мультимножину $W_s^{tw(kw)} = \{ n_i^{st}(w_i) \mid w_i \in TW^{kw} \}$, де n_i^{st} - кількість появ лексеми w_i в повідомленнях аналізованого масиву. Розглянемо об'єднання всіх повідомлень кожного окремого користувача usr_j як цілісні інформаційні об'єкти $tw_j^{usr(kw)} = \{ tw_i \mid usr(tw_i) = usr_j \}$. Масив повідомлень розглянемо як об'єднання повідомлень окремих користувачів, тобто інформаційних об'єктів $tw_j^{usr(kw)}$ $TW_s^{usr(kw)} = \{ tw_j^{usr(kw)} \}$. Введемо узагальнене поняття семантичного поля. Під семантичним полем будемо розуміти деяку підмножину словника, елементи якої об'єднані деяким спільним семантичним поняттям. В загальному випадку такі поняття можуть об'єднувати ключові слова, які відносяться до підрозділів аналізованої тематики, а також об'єднувати групи користувачів із спільними інтересами. Уведемо множину лексем, в яку входять ключові слова - теги, імена користувачів та назви тематичних груп $Keywords = \{ keyword_i \}$. Використовуючи теорію аналізу формальних концептів [2,3,4] розглянемо формальний контекст як трійку

$$K_{usr}^{tw(kw)} = (TW_s^{usr(kw)}, Keywords, I_s),$$

де I_s - відношення $I_s \subseteq TW_s^{usr(kw)} \times Keywords$, яке описує зв'язки повідомлень користувачів із тегами в цих повідомленнях. Вважаємо, що $(tw_i^{usr(kw)}, keyword_j) \in I_s$, якщо тег $keyword_j$ зустрічається в масиві повідомлень $tw_i^{usr(kw)}$ деяку кількість раз. Відношення I_s можна розглядати як мультимножину

$$I_s = \{ n_{ij}^{usr}(tw_i^{usr}, keyword_j) \mid keyword_j \in tw_i^{usr(kw)}, n_{ij}^{usr} > n_{th}^{usr} \}.$$

Введення порогового значення n_{th}^{usr} є необхідними для того, щоб включити в розгляд лише теги понять, які активно обговорюються. Уведемо ґратку семантичних концептів. Для деяких $Ext \subseteq TW_s^{usr(kw)}$, $Int \subseteq Keywords$ визначимо такі відображення

$$Ext' = \{ keyword_j \in Keywords \mid tw_i^{usr(kw)} \in Ext : (tw_i^{usr(kw)}, keyword_j) \in I_s \}$$

$$Int' = \{ tw_i^{usr(kw)} \in TW_s^{usr(kw)} \mid keyword_j \in Int : (tw_i^{usr(kw)}, keyword_j) \in I_s \}$$

Множина Ext' описує ключові теги, які властиві документам множини Ext , а множина Int' описує документи, які володіють тегами множини Int . Уведемо семантичний концепт як пару $Concept = (Ext, Int)$, до якої належать повідомлення з множини $Ext \subseteq TW_s^{usr(kw)}$ та ключові теги з множини $Int \subseteq Keywords$ з такими умовами: $Ext' = Int, Int' = Ext$. Множину Ext називають об'ємом, а Int – змістом семантичного концепту $Concept$. Таким чином, об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх блогах теги, які утворюють множину змісту цього концепту. В семантичному контексті $K_{usr}^{tw(kw)}$ утворюється частково-впорядкована множина семантичних концептів: $\Psi(TW_s^{usr(kw)}, Keywords, I_s) = \{ Concept_m = (Ext_m, Int_m) \}$. Ґратку концептів часто відображають за допомогою діаграм Гассе. В аналізі семантичного контексту $K_{usr}^{tw(kw)}$ кожний елемент діаграми представляє семантичний концепт. Такі діаграми відображають внутрішню семантичну структурну організацію повідомлень користувачів та відповідних їм груп ключових тегів. Розглянемо поняття порядкового ідеалу та фільтра для деякої частково впорядкованої множини $(P, \leq)$. Порядковим ідеалом називають підмножину $J \subseteq P$ для якої $\forall x \in J, y \leq x \Rightarrow y \in J$. Порядковим фільтром називають підмножину $F \subseteq P$ для якої $\forall x \in F, y \geq x \Rightarrow y \in F$. Використання понять порядкового ідеалу та фільтра може бути ефективним в аналізі ґратки семантичних концептів. Порядковим ідеалом деякого концепта будуть концепти, які пов'язані з ним на діаграмі Гассе і знаходяться нижче нього включаючи концепт, який відповідає інфімуму ґратки. Порядковим фільтром деякого концепту є множина пов'язаних із ним концептів, які знаходяться вище нього в ґратці, включаючи концепт, який відповідає супремуму ґратки. Зміст деякого концепту є підмножиною змістів концептів, які належать до його порядкового ідеалу. З іншої сторони, об'єднання змістів концептів, які утворюють порядковий фільтр деякого концепту утворює зміст цього концепту. Інформативним для аналізу є також розгляд об'єднання порядкового фільтра та ідеалу. Множина змістів такого об'єднання утворює деяке семантичне поле, яке відображає множину взаємопов'язаних понять. В одній ґратці може знаходитись декілька таких незалежних об'єднань порядкових ідеалів та фільтрів.

На основі розрахованої ґратки семантичних концептів можна виявити асоціативні правила, які відображають семантичні структурні зв'язки між ключовими словами. Під асоціативним правилом контексту $K_{usr}^{tw(kw)} = (TW_s^{usr(kw)}, Keywords, I_s)$ будемо розуміти вираз

$$A \rightarrow B, A, B \subseteq Keywords$$

Підмножину A називають передумовою, а B – наслідком асоціативного правила $A \rightarrow B$. Важливими характеристиками асоціативних правил є підтримка (support) $Supp_{A \rightarrow B}$ та достовірність (confidence) $Conf_{A \rightarrow B}$, які можна обрахувати за такими виразами:

$$Supp_{A \rightarrow B} = \frac{|(A \cup B)'|}{|TW_s^{(kw)}|} \quad Conf_{A \rightarrow B} = \frac{|(A \cup B)'|}{|A'|}$$

Значення $Supp_{A \rightarrow B}$ характеризує частку повідомлень $TW_s^{(kw)}$, яка містить ознаки $A \cup B$. Величина $Conf_{A \rightarrow B}$ характеризує частку повідомлень із ключовими словами множини A , яка також містить ключові слова множини B . Актуальними для аналізу є правила із деяким заданим мінімальним значенням підтримки та достовірності: $Supp_{A \rightarrow B} > Supp_{min}$, $Conf_{A \rightarrow B} > Conf_{min}$.

Для реалізації експериментальних досліджень розроблено пакет прикладних програм на мові Perl. За допомогою цього пакету використовуючи API системи Twitter завантажено тестовий масив повідомлень, які містять ключове слово “software” а також хеш-тег “#software”. Тобто, відібрано повідомлення заданого тематичного напрямку пов’язаного із програмним забезпеченням. В загальному завантажено близько 70 тисяч твітів. Далі був сформований контекст, в якому рядки відображали об’єднані повідомлення кожного із дописувачів мікроблогу. Були відфільтровані стоп-слова. З розгляду були вилучені лексеми які користувач використовував менше 25 раз. Був сформований контекст, який містив 112 об’єктів та 99 атрибутів. На основі цього контексту отримано асоціативні правила, приклади яких наведено в таблиці 1.

Таблиця 1.

Приклади виявлених асоціативних правил.

№	Передумова A	Наслідок B	$Supp_{A \rightarrow B}$	$Conf_{A \rightarrow B}$
1	{android}	{phone, windows}	0.89%	50.0%
2	{security}	{internet}	1.78%	100.0%
3	{iphone}	{application, developer, technologies}	0.89%	50.0%
4	{stream, traffic}	{internet}	1.78%	100.0%
5	{developer, iphone}	{application, technologies}	0.89%	100.0%

Асоціативні правила, для яких $Conf_{A \rightarrow B} = 100\%$ утворюють імплікації, тобто є справедливими для всіх випадків появи передумови правила. Зокрема, в аналізованому масиві повідомлень були отримані такі імплікації: {engineering} $\Rightarrow$ {computer, hardware}, {#ecommerce} $\Rightarrow$ {business, shopping}, {developer, iphone} $\Rightarrow$ {application, technologies}.

Фрагмент розрахованої діаграми Гассе для отриманої ґратки семантичних концептів із виділеним порядковим ідеалом та фільтром концепту {internet} наведено на рис. 1. Семантична структура концептів на рис.1 відображає поняття та тематики, якими інтенсивно цікавиться деяка група користувачів, кожен із яких згадає наведені поняття не менше 25 разів у своїх повідомленнях. Очевидно, що ця семантична структура може бути не відображена в аналізі масиву твітів без групування за користувачами, оскільки деякі взаємопов’язані поняття можуть знаходитись в різних повідомленнях одного і того ж користувача, а отже не попадуть в часті множини лексем масиву повідомлень без групування.

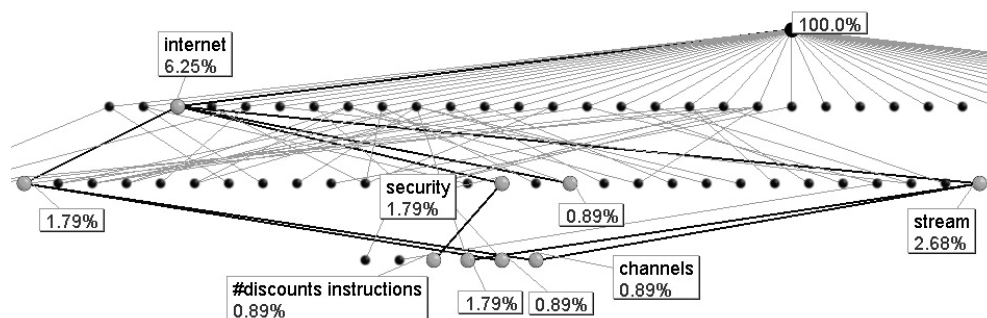


Рис.1 Порядковий ідеал та фільтр концепту {internet}.

Отже, в роботі запропонована модель ґратки семантичних концептів для аналізу тегів в повідомленнях згрупованих за користувачами мікроблогів. Введення поняття семантичного поля як множини тематично об'єднаних лексем зменшує обсяг необхідних обчислень внаслідок фільтрації масиву повідомлень. Фільтрація полягає у відкиданні лексем повідомлень, які не входять у задану семантичним полем тематику. Використання аналізу формальних концептів дає можливість утворити алгебраїчну ґратку семантичних концептів, характеристиками яких є об'єм та зміст. Об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх блогах групи тегів, які утворюють множину змісту цього концепту. Групування тегів користувачів здійснюється на основі множин змістів семантичних концептів. Утворена ґратка семантичних концептів дає можливість виявляти асоціативні правила у групах тегів, які є складовими змістів семантичних концептів. Ці правила відображають зв'язки між тегами, які характеризують семантичні поняття у повідомленнях користувачів.

Література

1. Брасегян А.А. 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.—512с.
2. Ganter B. Formal Concept Analysis: Mathematical Foundations/ B.Ganter, R.Wille.—Springer, 1999.
3. Kuznetsov S.O. Comparing Performance of Algorithms for Generating Concept Lattices / S.O. Kuznetsov, S.A. Obiedkov //Journal of Experimental and Theoretical Artificial Intelligence.—2002.—vol.14.—pp.189-216.
4. Cimiano P. Learning Concept Hierarchies from Text Corpora, using Formal Concept Analysis / P. Cimiano, A. Hotho, S. Staab //Journal of Artificial Intelligence Research.—2005.—vol.24.—pp.305-339.
5. Павлишенко Б.М. Групування текстових даних на основі моделі семантичного контексту / Б. М. Павлишенко // Східно-Європейський журнал передових технологій. — 2011. — №5/2(53). — С. 39–42.