УДК 519.765:519.767:004.93

Б.М. Павлишенко
Львівський національний університет імені Івана Франка

ГРУПУВАННЯ ТЕГІВ КОРИСТУВАЧІВ МІКРОБЛОГІВ НА ОСНОВІ РЕШІТКИ СЕМАНТИЧНИХ КОНЦЕПТІВ

Запропоновано модель решітки семантичних концептів для аналізу тегів у повідомленнях, згрупованих за користувачами мікроблогів. Показано, що використання цієї моделі є ефективним під час аналізу груп ключових слів та виявлення асоціативних правил в повідомленнях мікроблогів.

Ключові слова: інтелектуальний аналіз даних, аналіз формальних концептів, мікроблоги, семантичні поля.

The Grouping of Tags of Microblogs Users On The Basis of Semantic Concepts Lattice. The model of the semantic concept lattice for the analysis of tags in the messages grouped by microblogs users has been proposed in this work. It is shown that the use of this model is effective for the analysis of key words groups and for the detection of associative rules in the microblogs messages.

Key words: data mining, formal concepts analysis, FCA, microblogs, semantic fields.

Вступ

Інтелектуальний аналіз веб-контенту є складовою частиною сучасних інформаційних технологій. Система мікроблогів Twitter є одним із популярних засобів взаємодії користувачів за допомогою коротких повідомлень (не більше ніж 140 символів). Формат таких повідомлень є надзвичайно простий і дозволяє згадувати в тексті інших користувачів (наприклад, @username) та тематичні групи за допомогою хештегів з позначкою # (наприклад, #software). Повідомлення одночасно надсилаються згаданим в них користувачам та тематичним групам. Такий формат дає можливість за деяким ключовим словом виявляти повідомлення, які містять це слово, а також виявляти користувачів та групи, які мають відношення до тематики, заданої цим ключовим словом. Такі повідомлення також несуть інформацію про взаємозв'язок між окремими користувачами та ключовими словами – тегами. Лаконічність Twitter-повідомлень зумовлює високу густину тематично значимих ключових слів і створює перспективність досліджень мікроблогів засобами інтелектуального аналізу. Під тегами користувача розумітимемо такі ключові слова, які користувач часто вживає у своїх повідомленнях.

Аналіз останніх досліджень та публікацій.

Інтелектуальний аналіз слабко структурованих даних є ефективним методом дослідження текстових масивів [1, 2]. У такому аналізі використовують, зокрема, алгоритми пошуку частих множин ознак та асоціативних правил, за допомогою яких можна виявити взаємозв`язок між підмножинами даних [3–6]. Одним із перспективних напрямків аналізу даних є теорія аналізу формальних концептів [2–5]. У цій теорії розглядається відношення об'єктів та їх атрибутів, на основі якого будують алгебраїчну решітку 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. У роботі [8] використовують теорію аналізу формальних концептів для аналізу американських політичних блогів. Актуальним сьогодні 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.

Постановка задачі

Розглянемо теоретико-множинну модель, яка описує повідомлення мікроблогів, згрупованих за користувачами. У дослідженнях використаємо аналіз формальних концептів, який базується на теорії алгебраїчних решіток [1–4]. Проаналізуємо утворення семантичних концептів та асоціативних правил для груп тегів користувачів. На основі розглянутої теоретичної моделі проаналізуємо тестовий масив повідомлень системи мікроблогів Twitter.

Виклад основного матеріалу. Теоретична модель

Розглянемо модель, яка описує повідомлення мікроблогів, їх словник, користувачів та тематичні групи. Нехай вибрано деяке ключове слово *kw*, яке задає тематику повідомлень і є наявне у всіх повідомленнях, наприклад *kw=software*. Визначимо множину повідомлень мікроблогів:

$$TW^{kw} = \{\ tw_i \mid kw \in tw_i\ \}. \tag{1}$$

Загальний словник аналізованого масиву повідомлень розглянемо як мультимножину

$$W_s^{tw(kw)} = \{\ n_i^{st}(w_i) \mid w_i \in TW^{kw}\ \} \tag{2}$$

де n_i^{st} – кількість появ лексеми w_i в повідомленнях аналізованого масиву. Множину користувачів позначимо

$$USR = \{\ usr_i\ \}. \tag{3}$$

Розглянемо об'єднання всіх повідомлень кожного окремого користувача usr_j як цілісні інформаційні об'єкти

$$tw_j^{usr(kw)} = \{\ tw_i \mid usr(tw_i) = usr_j\ \}. \tag{4}$$

Масив повідомлень розглянемо як об'єднання повідомлень окремих користувачів, тобто інформаційних об'єктів $tw_j^{usr(kw)}$

$$TW_s^{usr(kw)} = \{\ tw_j^{usr(kw)}\ \}. \tag{5}$$

Оскільки всі повідомлення містять наперед задане ключове слово (в наших дослідженнях це слово – "*software*"), то такий масив повідомлень буде охоплювати деякий наперед заданий семантичний спектр інформації. Введемо узагальнене поняття семантичного поля [5]. Під семантичним полем розумітимемо деяку підмножину словника, елементи якої об'єднані деяким спільним семантичним поняттям. У загальному випадку такі поняття можуть об'єднувати ключові слова, які належать 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

$$Keywords = \{\ keyword_i\ \}\ . \tag{6}$$

Використовуючи теорію аналізу формальних концептів [1–4], розглянемо формальний контекст як трійку

$$K_{usr}^{tw(kw)} = \left(TW_s^{usr(kw)}, Keywords, I_s\right) \tag{7}$$

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 разів. Відношення I_s можна розглядати як мультимножину

$$I_s = \{\ n_{ij}^{usr}(tw_i^{usr}, keyword_j) \mid keyword_j \in tw_i^{usr(kw)}, n_{ij}^{usr} > n_{th}^{usr}\ \}. \tag{8}$$

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

$$Ext' = \{\ keyword_j \in Keywords \mid tw_i^{usr(kw)} \in Ext : (tw_i^{usr(kw)}, keyword_j) \in I_s\ \} \tag{9}$$

$$Int' = \{tw_i^{usr(kw)} \in TW_s^{usr(kw)} \mid keyword_j \in Int : (tw_i^{usr(kw)}, keyword_j) \in I_s\ \} \tag{10}$$

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

$$Concept = (Ext, Int) \tag{11}$$

до якої належать повідомлення з множини $Ext \subseteq TW_s^{usr(kw)}$ та ключові теги з множини $Int \subseteq Keywords$ з такими умовами:

$$\begin{cases} Ext' = Int, \\ Int' = Ext. \end{cases} \tag{12}$$

Множину Ext назвемо об'ємом, а Int – змістом семантичного концепту $Concept$. Отже, об'єм семантичного концепту об'єднує користувачів, які часто вживають у своїх блогах теги, які утворюють множину змісту цього концепту. У семантичному контексті $K_{usr}^{tw(kw)}$ утворюється частково-впорядкована множина семантичних концептів

$$\Psi(TW_s^{usr(kw)}, Keywords, I_s) = \{ \ Concept \ _m = (Ext_m, \ Int_m) \}, \tag{13}$$

Семантичний концепт

$$Concept \ _1 = (Ext_1, Int_1) \tag{14}$$

є менш загальним за об'ємом ніж концепт

$$Concept_2 = (Ext_2, Cont_2) \tag{15}$$

тобто виконується умова

$$(Ext_1, Int_1) \le (Ext_2, Int_2), \tag{16}$$

якщо

$$Ext_1 \subseteq Ext_2 \ \Leftrightarrow \ Int_1 \supseteq Int_2. \tag{17}$$

У цьому випадку концепт $Concept_2$ можна вважати узагальненням концепту $Concept_1$. Семантичний концепт можна розглядати як підматрицю семантичного контексту, яка повністю заповнена одиницями. Решітку 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 $(P, \le)$. Порядковим ідеалом називають підмножину $J \subseteq P$, для якої

$$\forall x \in J, y \le x \Rightarrow y \in J. \tag{18}$$

Порядковим фільтром називають підмножину $F \subseteq P$, для якої

$$\forall x \in F, y \ge x \Rightarrow y \in F. \tag{19}$$

Використання понять порядкового ідеала та фільтра може бути ефективним в аналізі решітки семантичних концептів. Порядковим ідеалом деякого концепта будуть концепти, які пов'язані з ним на діаграмі Гассе і розташовані нижче від нього, а також і концепт, який відповідає інфімуму решітки. Порядковим фільтром деякого концепту є множина пов'язаних із ним концептів, які розташовані вище від нього в решітці, а також і концепт, який відповідає супремуму решітки. Зміст деякого концепту є підмножиною змістів концептів, які належать до його порядкового ідеалу. З іншого боку, 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 ідеала. Множина змістів такого об'єднання утворює деяке семантичне поле, яке відображає множину взаємопов'язаних понять. В одній решітці може перебувати декілька таких незалежних об'єднань порядкових ідеалів та фільтрів.

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

$$A \to B, \ \ A, B \subseteq Keywords \tag{20}$$

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

$$Supp_{A\to B} = \frac{\left|(A \cup B)'\right|}{\left|TW_S^{(kw)}\right|} \tag{21}$$

$$Conf_{A\to B} = \frac{\left|(A \cup B)'\right|}{\left|A'\right|} \tag{22}$$

У випадку коли $Conf_{A\to B} = 1$ асоціативне правило (21) є імплікацією, тобто виконується завжди, коли зустрічається передумова A. Значення $Supp_{A\to B}$ характеризує частку повідомлень $TW_S^{(kw)}$, яка містить ознаки $A \cup B$. Величина $Conf_{A\to B}$ 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:

$$Supp_{A\to B} > Supp_{\min} \tag{24}$$

$$Conf_{A\to B} > Conf_{\min} \tag{25}$$

Експериментальна частина

Для реалізації експериментальних досліджень розроблено пакет прикладних програм мовою Perl. За допомогою цього пакета, використовуючи API системи Twitter, завантажено тестовий масив повідомлень, які містять ключове слово “software”, а також хеш-тег “#software”. Тобто, відібрано повідомлення заданого тематичного напряму, 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 ніж 25 разів. Був сформований контекст, який містив 112 об’єктів та 99 атрибутів. На основі цього контексту отримано асоціативні правила при підтримці більше 0.1%, приклади цих правил наведено в таблиці.

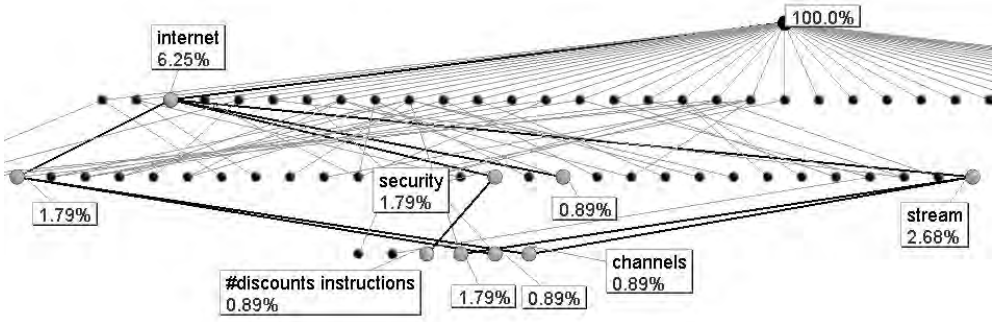
Приклади виявлених асоціативних правил

№	Передумова A	Наслідок B	$Supp_{A\to B}$	$Conf_{A\to B}$
1	{android}	{phone, windows}	0.89%	50.0%
2	{windows}	{office}	0.89%	50.0%
3	{#linux}	{#opensource}	0.89%	50.0%
4	{#income}	{programmer}	0.89%	100.0%
5	{engineering}	{computer, hardware}	0.89%	100.0%
6	{iphone}	{application, developer, technologies}	0.89%	50.0%
7	{#ecommerce}	{business, shopping}	0.89%	100.0%
8	{security}	{internet}	1.78%	100.0%
9	{freeware}	{download}	0.89%	100.0%
10	{stream, traffic}	{internet}	1.78%	100.0%
11	{developer, iphone}	{application, technologies}	0.89%	100.0%
12	{internet, online}	{#movies, #tv, satellite}	0.89%	100.0%

Асоціативні правила, для яких $Conf_{A\to B} = 100\%$ утворюють імплікації, тобто є справедливі для всіх випадків появи передумови правила. Для наведених у таблиці прикладів можна виявити, зокрема, такі імплікації:

{internet, online} => {#movies, #tv, satellite},
{developer, iphone}=>{application, technologies},
{freeware}=>{download},
{#ecommerce}=>{business, shopping}.

Потрібно відмітити, що ці асоціативні правила є імплікаціями лише для відфільтрованої за заданими умовами вибірки повідомлень мікроблогів і в загальному випадку можуть не бути такими. Фрагмент розрахованої діаграми Гассе для отриманої решітки семантичних концептів із виділеним порядковим ідеалом та фільтром концепту {internet} наведено на рисунку.



Порядковий ідеал та фільтр концепту {internet}

Семантична структура концептів на рисунку, відображає поняття та тематики, якими інтенсивно цікавиться деяка група користувачів, кожен із яких згадує наведені поняття не менше ніж 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 перебувати в різних повідомленнях одного і того ж користувача, а отже, не попадуть в часті множини лексем масиву повідомлень без групування.

Висновки

Запропонована модель решітки семантичних концептів для аналізу тегів у 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 решітку 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 решітка 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.

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. – 512 с. 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. El Qadi Abderrahim Formal Concept Analysis for Information Retrieval / Abderrahim El Qadi, Driss Aboutajedine, Yassine Ennouary // (IJCSIS) International Journal of Computer Science and Information Security. – 2010. – vol. 7, N. 2. 6. Agrawal R. Fast algorithm for mining association rules / R. Agrawal, R. Srikant // Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pages 487–499, Santiago, Chile, Sept. 1994. 7. Левицкий В.В. Экспериментальные методы в семасиологии / В.В. Левицкий, И.А. Стернин. – Воронеж: Изд-во ВГУ, 1989. – 192 с. 8. Klimushkin M. Formal Concept Analysis of the US Blogosphere during the 2008 Presidential Campaign / M. Klimushkin, D. Chetvericov, A. Novokreshchenova // Academic papers of the 8th international session of the HSE "Baltic practice", 2009. – p. 151–160.