УДК 004.912 : 81'33

О. Ю. Стадній

АВТОМАТИЗОВАНА КЛАСИФІКАЦІЯ УЧАСНИКІВ СОЦІАЛЬНИХ МЕРЕЖ НА ОСНОВІ ЛІНГВІСТИЧНОГО АНАЛІЗУ МІКРОБЛОГІВ

Вінницький національний технічний університет

Анотація

У роботі запропоновано підхід до автоматизованої класифікації учасників соціальних мереж на основі лінгвістичного аналізу мікроблогів. Розглянуті основні поняття предметної області, обґрунтовано вибір методів та інструментальних засобів для реалізації підходу.

Ключові слова: класифікація користувачів, соціальні мережі, мікроблоги, тезаурус, база знань, комп'ютерна лінгвістика.

Abstract

In this work, instruments of social network automated classification on the base of linguistic analysis of microblogs were explored. Basic points of subject area were overviewed, reasons of choice the methods and tools for practical realization of approach were examined.

Keywords: classification of users, social networks, microblogs, thesaurus, knowledge base.

Сучасне життя неможливо уявити без мережі Інтернет та, зокрема, соціальних мереж. Вони проникли у наше повсякдення і не часто можна зустріти людину, яка не має акаунта на Facebook, Instagram, Twitter чи забороненому Vkontakte – а то і на всіх одночасно. Саме завдяки таким ресурсам люди оперативно отримують та поширюють інформацію, спілкуються. Але сучасні соцмережі – це не лише місце, де можна поспілкуватись зі знайомими онлайн. Вони є хорошим майданчиком для маркетингу, соціологічних досліджень, знайомств, об'єднань груп за інтересами, інформування, впливу, пропаганди, інструментом спецслужб по виявленню соціально небезпечних елементів тощо.

Алгоритми класифікації користувачів вже зараз широко використовуються самими соцмережами для підтримки зацікавленості користувачів. Відповідні методи зазвичай будуються на основі бази знань (БЗ), яка, в свою чергу, будується на основі тезаурусу. Сучасні БЗ працюють сумісно з системами пошуку та виведення інформації [1]. Бази знань в різних країнах вивчали такі науковці, як Ф. Хейс-Рот, Д. Ватерман, Д. Ленарт, К. Грін, Д. Лакхем, Р. Бальзер, Т. Чітхем, К. Річ, М. Кауфман, Е. Файгенбаум, М. Ярке, С. Крішна, Т. Гаврилова, В. Хорошевський, Н. Гулякіна, С. Субботін, С. Шереметьєва, С. Пасмурнов, О. Фіртич, М. Іванов, П. Осмінін.

Бази знань містять низки понять, сутностей і зв'язків. Відомі зразки БЗ включають DBLP, Google Scholar, Internet Movie Database, YAGO, DBpedia, Wolfram Alpha, Freebase. Пошукові сайти, як-от Google чи Bing застосовують БЗ для того, щоб найкращим чином відповісти на запити користувачів. Так само вчиняють amazon.com і walmart.com. Іншим прикладом є Siri – голосовий асистент від iPhone, що використовує БЗ для обробки запитань (запитів користувачів) і генерації відповідей на них. Echonest.com вибудовує велику базу знань про музику, яку потім використовує для запуску низки програм, таких як рекомендації, плейлісти і аудіоаналіз. Інші зразки включають використання БЗ для знаходження експертів в біомедицині, для аналізу соціальних медіа, пошуку в Deep Web і «добування» соціальних даних [1].

Для розв'язання задачі автоматизованої класифікації учасників соціальних мереж на основі лінгвістичного аналізу мікроблогів пропонується розробити метод побудови та автоматичного розширення бази знань предметної області мікроблогів на основі програмно доступного тезаурусу. Для цього необхідно забезпечити автоматизоване визначення термінів тезаурусу та взаємозв'язків між ними з тексту мікроблогу, побудувати механізм самовдосконалення тезаурусу та визначити простір значимих параметрів для класифікації авторів – учасників соціальних мереж.

Розглянемо основні поняття предметної області. Будемо вважати, що онтологія – це система, що складається з набору понять і тверджень про ці поняття, на основі яких можна описувати класи, відносини, функції та індивіди. Наприклад, якщо ми розглядаємо просту предметну галузь, що

описує кубики на столі, то онтологія – це набір можливих положень кубиків, а не конкретне їхнє розташування в поточний момент часу.

Існує два альтернативних підходи до створення і дослідження онтологій: перший (формальний) заснований на логіці (предикатів першого порядку, дескриптивній, модальній тощо); другий (лінгвістичний) ґрунтується на вивченні природної мови (зокрема, семантики) і побудові онтологій на великих текстових масивах, так званих корпусах [2].

У даний час ці підходи тісно взаємодіють. Відбувається пошук зв'язків, що дозволяють комбінувати відповідні методи. Тому іноді буває складно відокремити лексичні онтології з елементами формальної аксіоматики від логічних систем з використанням лінгвістичних знань.

Тезауруси є специфікаціями онтологій. Тезаурус – це словник, набір відомостей, корпус або зведена інформація, що повномірно охоплює поняття, визначення і терміни спеціальної галузі знань або сфери діяльності, що має сприяти правильній лексичній, корпоративній комунікації (розуміння в спілкуванні і взаємодії осіб, пов'язаних однією дисципліною чи професійними обов'язками); в сучасній лінгвістиці – особливий різновид словників, в яких вказані семантичні відносини (синоніми, антоніми, пароніми, гіпоніми, гіпероніми тощо) між лексичними одиницями. Тезауруси є одним з дієвих інструментів для опису окремих предметних областей і наразі широко використовуються у сферах, які обслуговуються комп'ютерною лінгвістикою [3].

Тезауруси містять додаткову семантику, визначаючи зв'язки між термінами. Відношеннями, властивими для тезаурусів є синонімія, ієрархічне відношення і асоціація. Ранні ієрархії термінів, що з'явилися в мережі, визначали терміни через операції узагальнення і уточнення. Yahoo, наприклад, ввела невелике число категорій верхнього рівня, таких, як "предмети одягу". Потім "плаття" визначалося як вид (жіночого) одягу. Явна ієрархія Yahoo не відповідала в точності формальним властивостям ієрархічних відносин Підклас-Клас. У таких ієрархіях може зустрітися ситуація, в якій екземпляр класу-нащадка не є екземпляром класу-пращура. Наприклад, загальна категорія "предмети одягу" має підкатегорію "жіночі" (яка повинна була б більш точно називатися "жіночі предмети одягу"), а ця категорія, у свою чергу, охоплює підкатегорії "аксесуари" і "сукні". Ясно, що аксесуари, наприклад "брошки", не є предметами одягу. Тут не виконується важлива властивість відносин Підклас-Клас – транзитивність.

Основним відношенням (зв'язком) між термінами в тезаурусі є зв'язок між більш широкими (більш виразними) і більш вузькими (більш спеціалізованими) поняттями. Часто виділяють 2 підвиди цих відношень:

1. Один термін позначає поняття, що є частиною поняття, яке позначається іншим терміном (наприклад, «наука» і «математика», «математика» і «теорія чисел»)
2. Один термін, що називає елемент класу, позначається іншим терміном («гірські райони» і «Кавказ»).

Це відношення ґрунтується на множині термінів і є відношенням часткового порядку, тобто множина термінів з такими зв'язками утворює ациклічний граф або поліієрархічну структуру.

Існують також і інші зв'язки між термінами. Наприклад, одне поняття або концепція може бути позначено кількома термінами, які є синонімами. Деякі терміни можуть бути антонімами для інших. Часто серед термінів, які стосуються одного поняття, виділяють єдиний (для кожної мови тезауруса) найкращий (найбільш відповідний) термін, який найточніше характеризує або позначає дане поняття. Інші терміни є менш придатними.

Крім вищеописаних, між термінами можуть існувати також і інші, асоціативні зв'язки, якщо поняття, що позначаються цими термінами, як-небудь пов'язані між собою за смислом (сенсом) – за винятком описаних вищих ієрархічних зв'язків.

У багатомовних тезаурусах існують також зв'язки еквівалентності між термінами різними мовами. Виділяють повну (сувору) еквівалентність і кілька видів часткової (нестрогої) смислової еквівалентності термінів на різних мовах.

Тезаурус часто містить коментарі до термінів, що розкривають для користувача семантику терміна, а також пояснюють, як його слід використовувати.

Тезауруси застосовуються, перш за все, для класифікації та пошуку інформаційних ресурсів. При цьому в кожному ресурсі при класифікації можуть бути співставлені одне або більше понять, описаних термінами в тезаурусі, а користувач, який здійснює пошук, може по тезаурусу знайти цікаві для нього поняття в даній галузі, а також всі терміни, що їх характеризують. Отже, на основі зв'язків тезауруса відбувається розширення пошукового запиту (розширення слів запиту синонімічними,

більш загальними або більш приватними за змістом термінами). Навігація по зв'язках тезауруса допомагає чіткіше сформулювати сам запит.

Існує ряд стандартів різного рівня значущості і опрацьованості на формат представлення тезаурусів. Ці стандарти представляють тезаурус у вигляді набору об'єктів кількох типів, між якими може бути кілька типів зв'язків. Деякі стандарти (наприклад, стандарт ANSI / NISO Z39.19-1993) регламентують також формат уявлення тезауруса в лінеаризованому (текстовому) вигляді, придатному для сприйняття як машиною, так і людиною [4].

Основними документами, що регламентують формат уявлення тезауруса, є стандарти ISO 2788-1986 для опису одномовних тезаурусів, і ISO 5964-1985 для багатомовних.

Власне, на основі програмно доступного тезаурусу пропонується створити БЗ для поставленої задачі автоматизованої класифікації учасників соціальних мереж на основі лінгвістичного аналізу мікроблогів. База знань – це технологія, яка застосовується для зберігання складної структурованої та неструктурованої інформації, що використовується комп'ютерною системою. Початкове використання цього терміну було пов'язане з експертними системами, які були першими, що базуються на знаннях.

Первинне використання терміну БЗ полягало в описі однієї з двох підсистем системи, заснованої на знаннях. Така система складається з бази знань, що представляє факти про світ та механізм висновків, який може пояснити ці факти та правила використання та інші форми логіки для виявлення нових фактів або висвітлення невідповідностей. У статичній її частині зберігаються довгострокові знання, що описують розглянуту предметну область у вигляді загальних фактів (фраз без умов, що містять твердження, які завжди є абсолютно вірними) і правил (тверджень, істинність яких залежить від умов, що утворюють тіло правила), які описують доцільні перетворення фактів цієї області з метою створення нових фактів або гіпотез [5].

Для визначення простору значимих параметрів для класифікації авторів – учасників соціальних мереж з БЗ пропонується застосувати відомі методи машинного навчання. Дослідження вирішено провести на базі соціальної мережі Facebook через її найбільшу популярність у всьому світі та зручний власний інструмент для розробників – Graph API.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Omkar Deshpande Building, Maintaining, and Using Knowledge Bases: A Report from the Trenches / Omkar Deshpande, Digvijay S. Lamba, Michel Tourn, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, AnHai Doan.
2. Добров Б. В. Онтологии и тезаурусы: модели, инструменты, приложения / Добров Б. В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. — М.: Бином. Лаборатория знаний, 2009. — 173 с. — ISBN 978-5-9963-0007-5.
3. Большая советская энциклопедия под редакцией К. А. Штрома. – т. 25. – С. 60.
4. Nirenburg S., Raskin V. Ontological Semantics. – Cambridge, MA, 2004. – 265 p.
5. Krishna, S. Introduction to Database and Knowledge-based Systems. – Singapore: World Scientific Publishing, 1992. – ISBN 981-02-0619-4.

Стадній Олександра Юріївна, студентка групи 2СІ-14б, Вінницький національний технічний університет, м. Вінниця, e-mail: alix.stadniy@gmail.com

Науковий керівник: ***Бісікало Олег Володимирович*** – д-р техн. наук, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця

Stadnii Oleksandra Yuriyivna, student of the Faculty of Automation, Electronics and Computer Control Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: alix.stadniy@gmail.com

Supervisor: ***Bisikalo Oleg V.*** – Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia