УДК 004.773.2
DOI 10.31891/2307-5732-2020-285-3-11

С. С. ФЕДУШКО, Ю. О. СЄРОВ
Національний університет «Львівська політехніка»

ПРОЕКТУВАННЯ ТА РОЗРОБЛЕННЯ СПЕЦІАЛІЗОВАНОГО ЛІНГВІСТИЧНОГО СЛОВНИКА ДЛЯ ВИЗНАЧЕННЯ ПЕРСОНАЛЬНИХ ХАРАКТЕРИСТИК ВЕБ-КОРИСТУВАЧА

У статті представлено результати проектування та розроблення спеціалізованого словника для визначення персональних характеристик веб-користувача. У дослідженні розроблено метод дослідження інформаційних слідів веб-користувачів, який передбачає, на відміну від інших методів, використання спеціалізованого словника соціально-демографічних маркерів, що дало змогу сформувати систему лінгво-комунікативних індикаторів для верифікації даних користувачів віртуальних спільнот. У процесі верифікації веб-контенту спеціалізований словник соціально-демографічних маркерів виконує фундаментальну функцію, оскільки соціально-демографічні маркери є основною складовою комп'ютерно-лінгвістичного аналізу інформаційного сліду користувачів веб-спільноти.

Ключові слова: віртуальні спільнота, спеціалізований лінгвістичний словник, персональні характеристики, веб-користувач, інформаційна схема.

S. FEDUSHKO, YU. SYEROV
Lviv Polytechnic National University

DESIGN AND DEVELOPMENT OF A SPECIALIZED LINGUISTIC DICTIONARY TO DETERMINE THE PERSONAL CHARACTERISTICS OF A WEB USER

The study examines the results of designing and developing a specialized dictionary to determine the personal characteristics of a web user. The study developed a method of studying the information traces of web users, which involves the use of a specialized dictionary of socio-demographic markers. Using this method allows to form a system of linguistic and communicative indicators for verification of data of users of virtual communities. Scientific novelty of research results is a scientific justification and implementation of the first stage in the development of methods and means of computerized linguistic analysis of the reliability of the personal characteristics of users of communities is the design and development of specialized linguistic dictionary to determine the personal characteristics of web users. The purpose of the study is to develop a system of linguistic and communicative indicators based on a specialized dictionary of socio-demographic markers formed by analyzing the information traces of the educational sample of users of virtual communities. In this study, based on the developed method of research of information traces of web users, which provides, unlike other methods, will be designed and implemented for use a specialized dictionary of socio-demographic markers, which will form a system of linguistic and communicative indicators to verify data of users' web communities. Modelling of the information scheme of the dictionary is performed with the help of diagrammatic means of structural modelling. The means of forming personal portraits of users of virtual communities are built on the basis of modern web technologies. Socio-demographic characteristics are determined using markers contained in a specialized dictionary. The structure of the dictionary is developed according to the scheme of linguistic-communicative indicators of social and demographic characteristics. During verification, a specialized dictionary sociodemographic marker performs a fundamental function as socio-demographic markers are the main component of computerized linguistic analysis of the information track web user community. It is impossible to verify socio-demographic characteristics without a database of socio-demographic markers. Socio-demographic markers are a key link in the process of forming sets of linguistic-communicative indicators that indicate the affiliation of a web user to a certain socio-demographic characteristics.

Keywords: virtual community, specialized linguistic dictionary, personal characteristics, web user, information scheme.

Вступ. В час кризи COVID-19 спричиненої пандемією та світовим карантином практично уся соціальна та професійна діяльність людства перейшла у віртуальний світ, де є надлишок анонімності та високий рівень кібер-небезпеки. Розроблення дієвих методів аналізу веб-контенту, а саме, проектування та розроблення спеціалізованого словника для визначення персональних характеристик веб-користувача, є затребуваним у сьогочасних складних умовах в соціумі. Новизна результатів дослідження полягає у виконанні першого етапу у процесі розроблення методів та засобів комп'ютерно-лінгвістичного аналізу достовірності персональних характеристик користувачів веб-спільнот – проектування та розроблення спеціалізованого лінгвістичного словника для визначення персональних характеристик веб-користувача. Мета дослідження – розробити систему лінгво-комунікативних індикаторів на основі спеціалізованого словника соціально-демографічних маркерів сформованого шляхом аналізу інформаційних слідів навчальної вибірки користувачів віртуальних спільнот. У цьому дослідженні на основі розробленого методу дослідження інформаційних слідів веб-користувачів, який передбачає, на відміну від інших методів, буде спроектовано та запроваджено для використання спеціалізований словник соціально-демографічних маркерів, що дасть змогу сформувати систему лінгво-комунікативних індикаторів для верифікації даних користувачів віртуальних спільнот.

Аналіз досліджень та публікацій. Соціально-демографічна ідентичність веб-особистості користувача різноманітних соціальних комунікацій всесвітньої мережі перебуває в полі уваги багатьох суміжних наук: соціології, політології, психології, культурології, менеджменту, етнології, криміналістики, ораторського мистецтва, економіки тощо. Враховуючи це, з метою дослідження варіативних ідентичностей і самопрезентації у віртуальних спільнотах веб-користувача, віртуальну ідентифікацію слід аналізувати, застосовуючи міждисциплінарні підходи. Саме міждисциплінарний комплексний підхід використано для розроблення моделей та методів організації життєвого циклу віртуальної спільноти [1] та для моделювання

програмного комплексу перевірки особистих даних веб-користувачів [3]. Окремими важливими напрямками у досліджень є аналіз значення та складання словника [2], аналіз спеціалізованих словників [4], мультимодальної комунікації користувачів Інтернету [5], використання нечікої логіки для моделювання поведінки людини [6], аналіз лексикографічних внесок феміністичних словників [7]. Науковці активно досліджують мовні особливості емоційних лінгвістичних компонентів у дискурсі [8], семантичний аналіз для виявлення інформаційних та комунікаційних загроз користувачів онлайн-сервісів [9] та аналізують гендерні відмінності у онлайн-комунікації [10].

Розробка структури спеціалізованого словника для визначення персональних характеристик веб-користувача. Моделювання інформаційної схеми словника виконано за допомогою діаграмних засобів структурного моделювання. Засоби формування персональних портретів користувачів віртуальних спільнот побудовано на основі сучасних веб-технологій. Соціально-демографічні характеристики (СДХ) визначаємо за допомогою маркерів, які містяться у спеціалізованому словнику. Структуру словника розроблено відповідно до схеми лінгво-комунікативних індикаторів СДХ. Для моделювання інформаційної схеми обрано нотацію Баркера. Розглянемо інформаційну схему словника, що наведено на рис. 1, прокоментувавши сутності та їхні основні атрибути.

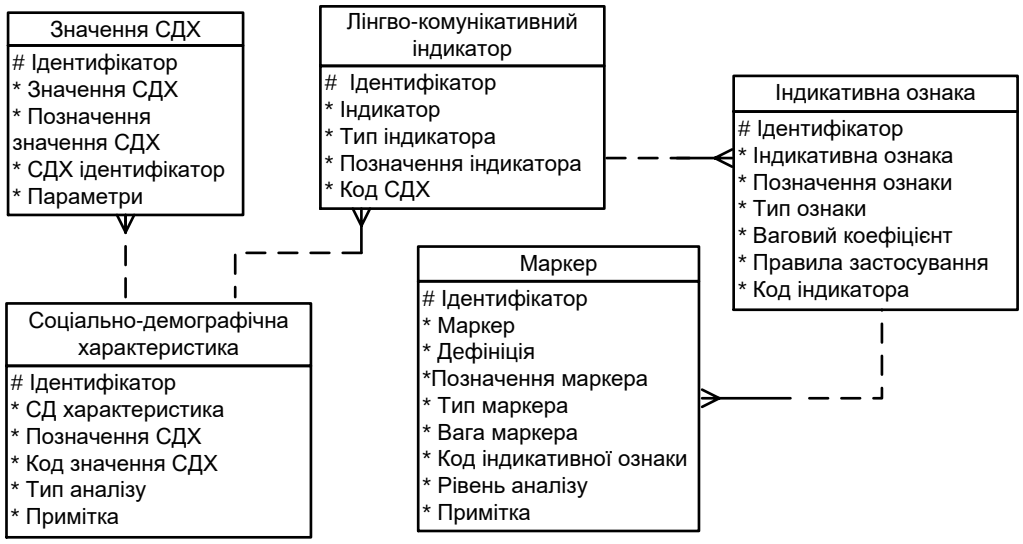

Рис. 1. Інформаційна схема словника маркерів персональних характеристик веб-користувача

Сутність «Соціально-демографічна характеристика» містить дані про соціально-демографічні характеристики, на основі яких будується соціально-демографічний портрет (profile) веб-користувача.

Залежно від типу інформаційного наповнення, створеного користувачем віртуальної спільноти, аналіз проводимо певним способом.

Атрибут «СД характеристика» містить найменування всіх досліджуваних СД характеристик, які є складовими СД портрету. Можливими значеннями цього атрибуту є: *Вік, Ґендер, Освіченість* та *Сфера діяльності.* Цим значенням атрибуту відповідають значення *Атрибуту «Позначення СДХ»*, який містить: *Age, Gend, Edu* та *Sphere*.

Атрибут «Тип аналізу» містить конкретні типи аналізу, наприклад, *«Лінгвістичний»*, якщо проводимо аналіз лексичних одиниць, та *«Графічний»* – у випадку наявності у інформаційному наповнені графічних елементів та фотоматеріалів.

Сутність "Значення СДХ" містить інформацію про значення, які присвоюються конкретній СДХ користувача ВС. Кожна СДХ має два значення. Тож відповідно, *Атрибут «Значення СДХ»* містить найменування значення СДХ відповідно до *Атрибута «СДХ ідентифікатор»*.

Так, СДХ атрибуту присвоюють одне з двох значень: *Чоловік /Жінка*, для СДХ вік – *Підліток / Дорослий*, для СДХ сфера діяльності – *Формальні науки/Природничі науки / Суспільні науки*, для СДХ рівень освіченості – *Високосвічений / Достатньоосвічений/Низькоосвічений*.

Для зручності дослідження та систематизації ознак кожному значенню присвоюється унікальне позначення. Напр., *M/F, Ado/Adu, FSc/LSc/SSc* та *HEdu/MEdu/LEdu*. Ці значення зберігаються у *Атрибуті «Позначення значення СДХ»*.

Сутність «Лінгво-комунікативний індикатор» містить інформацію за допомогою якої адміністратор ідентифікує аутентифікованого користувача віртуальної спільноти, групу користувачів та їх СД належність. Кожен лінгво-комунікативний набір містить лише унікальні індикатори.

Атрибут «Тип індикатора» містить набори лінгво-комунікативних індикаторів, які є результатом комп'ютерно-лінгвістичного аналізу інформаційних слідів конкретних користувачів веб-спільнот. Так, прикладом лінгво-комунікативного індикатора СДХ є:

– СДХ вік – *«Сленгова варіація», «Текстова економія», «Некодифіковані одиниці та невербальні засоби»* тощо;

– СДХ ґендер – *«Емоційна складова», «Спосіб вираження змісту», «Конкретизація»* тощо;

– СДХ рівень освіченості – *«Комбінації літер», «Комбінації символів», «Використання транслітерації»* тощо;

– СДХ сфера діяльності – «*Фізико-математична, технічна та економічна сфера», «Медична сфера», «Юридична сфера»* тощо.

Кожен індикатор має унікальне позначення, що міститься у *Атрибуті «Позначення індикатора»*. Прикладом таких позначень відповідно є: СДХ вік – *"AGE–B", "AGE–D"* та *"AGE–E"* тощо; СДХ ґендер – *"GENDER-A", "GENDER-F"* та *"GENDER-K"* тощо; СДХ рівень освіченості – *"EDU–A", "EDU–C"* та *"EDU–E"* тощо; СДХ сфера діяльності – *"SPHERE–A", "SPHERE–E"* та *"SPHERE–J"* тощо.

Сутність «Індикативна ознака» містить дані про набори індикативних ознак, які властиві веб-комунікації конкретного користувача віртуальної спільноти.

Атрибут «Тип ознаки» містить іменування типу індикативної ознаки. Так, наприклад, індикативні ознаки: *«Вказівка на особу, подію тощо», «Вказівка на особу, яка говорить про себе»* та *«Вказівка на кількох мовців»*, належать до лінгво-комунікативного індикатора ґендеру користувача ВС – *«Вказівка та інструкція»*.

Кожній індикативній ознаці присвоюється позначення, тобто певною міткою, для подальшої ідентифікації цих ознак. Ці дані містяться у *Атрибуті «Позначення ознаки»*. Наприклад, позначення індикативних ознак ЛКІ (лінгво-комунікативний індикатор) ґендеру користувачів веб-спільноти *«Вказівка та інструкція» (GENDER-D) є D(1.1), D(1.2)* та *D(1.3)*. *Атрибут «Ваговий коефіцієнт»* містить інформацію про коефіцієнт міри вираження конкретного маркера СДХ у інформаційному слідів користувачів веб-спільнот.

Атрибут «Правила застосування» містить інформацію про використання шаблонів регулярних виразів для виявлення СД маркерів у інформаційному слідів користувачів веб-спільнот.

Сутність «Маркер» містить дані про мовні та графічні ознаки, які вказують на належність користувача веб-спільноти до конкретної СДХ. Визначення маркерів інтернет-комунікації конкретного УВС здійснюється з допомогою аналізу у його інформаційного сліду.

Атрибут «Дефініція» містить стисле логічне визначення, яке містить у собі найістотніші ознаки конкретного маркеру. Наприклад, лінгвістичний маркер *«холівар» – одвічна суперечка на комп'ютерну тематику;* графічний маркер «:-[» *або* «  » – зашарітись, засоромитись, зніяковіти. *Атрибут «Позначення маркера»* містить унікальне позначення кожного маркеру. Ймовірні значення цього атрибуту: M(1), M(2), M(3)… M(N).

В контексті належності до відповідної індикативної ознаки відповідного лінгво-комунікативного індикатора відповідної СДХ виглядає наступним чином: *M(45)-AGE-B(1.2)*, тобто лінгвістичний маркер «*холівар*» індикативної ознаки *«Комп'ютерний слег»* лінгво-комунікативного індикатора *«Сленгова варіація»* значення соціально-демографічної характеристики *«Підліток»* СДХ *«Вік»; M(76)-GENDER-A(1.1)*, тобто графічний маркер ":-[" індикативної ознаки *«Згадка про емоції та почуття»* лінгво-комунікативного індикатора *«Емоційна складова»* значення соціально-демографічної характеристики *«Жінка»* СДХ *«Ґедер»*. *Атрибут «Тип маркера»* містить інформацію про типи маркерів. Так, цей атрибут може приймати такі значення, як: *«Лінгвістичний»* або *«Графічний»*.

Важливість маркера визначається його вагою. Вага маркера обраховується методом групового урахування аргументів. Значення ваги кожного маркеру міститься у *Атрибуті «Вага маркера»* містить.

Аналіз інформаційного сліду користувача веб-спільноти може здійснюватися тільки на певному рівні або на всіх можливих рівнях. Інформація про можливий рівень аналізу кожного маркера міститься у *Атрибуті «Рівень аналізу»*. Наприклад, відповідно до аналізу ІС маркери типу *«Лінгвістичний маркер»* можуть приймати такі значення: *«аналіз на рівні слів», «аналіз на рівні фраз»* та *«аналіз на рівні речень»*, а маркери типу *«Графічний маркер» – «аналіз аватару», «аналіз графічного підпису»* та *«аналіз графічних символів вираження емоцій»*.

Вибір та опрацювання персональних характеристик користувачів веб-спільноти, які є важливими для модерування спільнотою і потребують перевірки на достовірність, здійснюється згідно з формальною моделлю СД характеристик. Для функціонування компоненти необхідні дані, які є результатами роботи компонент:

1. Компонента формування наборів лінгво-комунікативних індикаторів.

2. Компонента формування інформаційного сліду.

При умові хоча б базового формування інформаційного сліду та наборів лінгво-комунікативних індикаторів компонента є логічним продовженням процесу перевірки достовірності персональних характеристик користувачів віртуальних спільнот.

Формування наборів лінгво-комунікативних індикаторів здійснюється на основі спеціалізованого словника соціально-демографічних маркерів, який постійно оновлюється.

Головними функціями компоненти валідації персональних характеристик є:

– виявлення недостовірної інформації у обліковому записі користувача віртуальної спільноти методом комп'ютерно-лінгвістичного аналізу інформаційного сліду цього користувача;

– формування звіту для модераторів віртуальної спільноти, у випадку недостатньої кількість даних, для здійснення верифікації одної чи декількох соціально-демографічних користувача віртуальної спільноти;

– повідомлення модератора про неповноту заповнення спеціалізованого словника соціально-демографічних маркерів та некоректне формування на його основі наборів лінгво-комунікативних індикаторів.

Результати роботи цієї компоненти є вхідними даними для компоненти побудови соціально-демографічного портрета.

Розроблення словника соціально-демографічних маркерів користувача веб-спільноти. У процесі верифікації спеціалізований словник соціально-демографічних маркерів виконує фундаментальну функцію, оскільки соціально-демографічні маркери є основною складовою комп'ютерно-лінгвістичного аналізу інформаційного сліду користувачів веб-спільноти. Отож, без наявності бази соціально-демографічних маркерів здійснення верифікації соціально-демографічних характеристик є неможливим. Саме соціально-демографічні маркери є ключовою ланкою у процесі формування наборів лінгво-комунікативних індикаторів, що вказують на належність веб-користувача до певної соціально-демографічної характеристики. Користувацький інтерфейс спеціалізованого словника СД маркерів веб-користувача представлено на рис. 2.

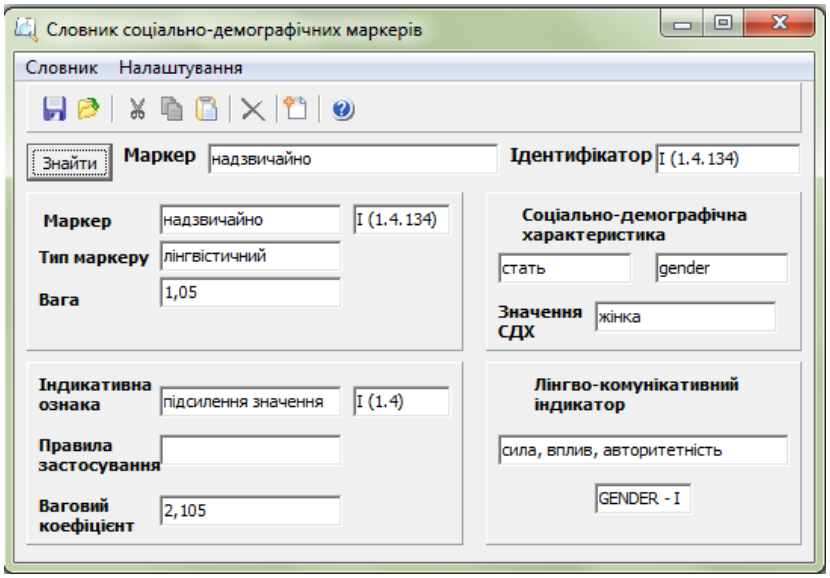

Рис. 2. Користувацький інтерфейс словника СД маркерів веб-користувача

Спеціалізований словник персональних характеристик користувачів віртуальних спільнот виконує функцію великої бази СД маркерів. Формування цієї бази відбувалось за таким алгоритмом:

– уніфікація, структурування та впорядкування за визначеним шаблоном величезного масиву веб-контенту;

– методом комп'ютерно-лінгвістичного аналізу та аналізу праць фахівців у різних галузях визначення маркерів користувачів веб-спільноти, що належить до конкретної соціально-демографічної характеристики;

– наповнення спеціалізованого інформаційного словника;

– в автоматизованому режимі системна адаптація бази словника відповідно до тематики, специфіки віртуальної спільноти та інтернет-комунікації веб-користувачів, які до неї належать;

– інтегрування словника у програмний засіб «Верифікатор персональних характеристик веб-користувача».

Вимоги до вмісту словника соціально-демографічних маркерів окреслено у його інформаційній моделі. Базуючись на даних цієї моделі та з врахування усіх потреб та специфіки веб-спільнот, розроблено інтерфейс словника СД маркерів, який необхідний для верифікації персональних характеристик веб-користувачів.

Висновки. Підтримка наповнення спеціалізованого лінгвістичного словника соціально-демографічних маркерів веб-користувача в актуальному стані та дотримання чіткої структури лінгво-комунікативних індикаторів є ключовими чинником у якості та надійності результатів верифікації персональних характеристик користувача віртуальної спільноти. Також від цих факторів залежать результати верифікації користувачів віртуальної спільноти. На основі результатів цього дослідження планується проектування та розроблення програмного засобу для верифікації персональних характеристик веб-користувача.

Література

1. Trach O. Development of Models and Methods of Virtual Community Life Cycle Organization / O. Trach, A. Peleshchyshyn // Advances in Artificial Systems for Medicine and Education II. AIMEE2018. Advances in Intelligent Systems and Computing, Springer, Cham. – 2018. – Vol. 902. – P. 473–483.

2. Nida Eugene A. Analysis of meaning and dictionary making / A. Nida Eugene // International Journal of American Linguistics. – 1958. – 24.4. – P. 279–292.

3. Fedushko S. Modeling software complex for web user personal data verification / S. Fedushko // Управління розвитком складних систем. – 2016. – 26. – P. 105–110.

4. Rice D. R. Corpus-based dictionaries for sentiment analysis of specialized vocabularies / D. R. Rice, C. Zorn // Political Science Research and Methods. – 2013. – P. 1–16.

5. Xie B. Multimodal computer-mediated communication and social support among older Chinese internet users / B. Xie // Journal of Computer-Mediated Communication. – 2008. – 13(3). – P. 728–750.

6. Ören T.I. Personality Representation Processable in Fuzzy Logic for Human Behavior Simulation / T.I. Ören, N. Ghasem-Aghaee // Proceedings of the 2003 Summer Computer Simulation Conference. – Montreal, PQ, Canada. – 2003. – P. 11–18.

7. Russell L. R. This is what a dictionary looks like: The lexicographical contributions of feminist dictionaries / L. R. Russell // International Journal of Lexicography. – 2012. – 25(1). – P. 1–29.

8. Amaghlobeli N. Linguistic features of typographic emoticons in SMS discourse / N. Amaghlobeli // Theory and Practice in Language Studies. – 2012. – 2(2). – P. 348.

9. Fedushko S. Semantic analysis for information and communication threats detection of online service users / S. Fedushko, E. Benova // Procedia Computer Science. – 2019. – 160. – P. 254–259.

10. Babal J. C. Linguistic analysis of pediatric residency personal statements: gender differences / J. C. Babal, A. D. Gower, J.G. Frohna, M. A. Moreno // BMC medical education. – 2019. – 19(1). – 392.

Надійшла / Paper received: 11.04.2020 Надрукована / Paper Printed : 04.06.2020