ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 3 (1), 2023

https://doi.org/10.23939/ictee2023.01.032

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ВЕБКОНТЕНТУ ДЛЯ ФОРМУВАННЯ СОЦІАЛЬНО-ЦИФРОВОЇ ІДЕНТИЧНОСТІ ВЕБКОРИСТУВАЧА

С. Федушко

Національний університет "Львівська політехніка", вул. С. Бандери, 12, Львів, 79013, Україна

Відповідальний за рукопис: С. Федушко (e-mail: solomiia.s.fedushko@lpnu.ua)

(Подано 26 березня 2023)

Висвітлено актуальну потребу в аналізі та розумінні поведінки вебкористувачів через безпрецедентну кількість цифрового контенту, що генерується та поширюється в інтернеті. Інтелектуальний аналіз вебконтенту представлений як ефективний спосіб вивчення та вилучення цінної інформації з онлайн-контенту, ураховуючи вебсайт, платформи соціальних мереж та інші цифрові джерела, щоб краще зрозуміти інтереси, вподобання та поведінку вебкористувачів. Здатність ідентифікувати вебкористувачів на основі їхньої поведінки в інтернеті важлива для таких галузей, як маркетинг, психологія та правоохоронна сфера. Однак з цим підходом пов'язані певні проблеми, такі як забезпечення конфіденційності та безпеки даних вебкористувачів, а також оцінка точності та надійності інструментів аналізу вебконтенту. Мета статті – огляд сучасного стану аналізу вебконтенту, його потенційних застосувань у різних галузях та ролі у формуванні цифрового майбутнього. У статті підкреслено важливість міждисциплінарного підходу до вивчення віртуальної ідентифікації та самопрезентації в онлайн-спільнотах з огляду на соціально-демографічні характеристики вебособистості, яка бере участь у соціальних взаємодіях. У статті також досліджено останні тенденції та розробки в галузі видобування вебданих, ураховуючи аналіз вебконтенту, аналіз вебструктури, аналіз використання вебсторінки та аналіз даних соціальних мереж. Запропоновано програмне рішення для здійснення інтелектуального аналізу вебконтенту із метою формування соціально-цифрової ідентичності вебкористувача за допомогою спеціалізованого словника маркерів контенту учасника вебспільноти.

Ключові слова: *вебконтент; видобування вебданих; соціальні мережі; інтернет; інтелектуальний аналіз; соціально-цифрова ідентичність.*
УДК 004.738.5+004.773.2

1. Вступ

Стрімкий розвиток соціальних мереж створив безпрецедентну кількість цифрового контенту, який генерується та поширюється онлайн. Оскільки мільярди людей отримують доступ до цього контенту і взаємодіють з ним, зростає потреба в аналізі та розумінні поведінки вебкористувачів. Одним із найефективніших способів отримати уявлення про характеристики соціально-цифрової ідентичності вебкористувача та його ймовірну поведінку і спрогнозувати його дії є аналіз вебконтенту [1, 2]. Аналіз вебконтенту – це процес вивчення та вилучення цінної інформації з інформаційного наповнення онлайн-спільнот, зокрема вебсайта, платформ соціальних мереж та інших цифрових джерел. Аналізуючи цей контент, дослідники можуть краще зрозуміти інтереси, вподобання та поведінку вебкористувачів [3–5]. Крім того, аналіз вебконтенту можна використовувати для ідентифікації вебкористувачів [6] на основі їхньої онлайн-активності та звичок споживання

контенту. Поєднання соціально-демографічних характеристик і цифрової поведінки та слідів, які є унікальними для людини в контексті її присутності та діяльності в інтернеті, формують її соціально-цифрову ідентичність.

Можливість ідентифікувати вебкористувачів на основі їхньої поведінки в інтернеті має важливе значення для різних галузей, зокрема маркетингу, психології, а також для правоохоронних органів. Наприклад, маркетологи можуть використовувати аналіз вебконтенту для створення цільових рекламних кампаній, а психологи – для вивчення зв'язку між поведінкою в інтернеті та психічним здоров'ям. У контексті правоохоронної діяльності аналіз вебконтенту допоможе ідентифікувати осіб, які можуть становити загрозу громадській безпеці. Попри потенційні переваги аналізу вебконтенту, з цим підходом пов'язані й певні виклики. Наприклад, забезпечення конфіденційності та безпеки даних вебкористувачів [7, 8] є основною проблемою, а точність і надійність інструментів аналізу вебконтенту необхідно ретельно оцінити. У цій статті здійснено огляд актуальності аналізу вебконтенту та ідентифікації вебкористувачів на основі аналізу їхнього вебконтенту. Мета статті – розглянути сучасний стан аналізу вебконтенту, обговорити проблеми, пов'язані з цим підходом, і дослідити потенційні можливості застосування цієї технології в різних галузях, що сприятиме кращому розумінню потенціалу та обмежень аналізу вебконтенту та його ролі у формуванні нашого цифрового майбутнього. Соціально-демографічні характеристики вебособистості [9–11], яка бере участь у соціальних взаємодіях у всесвітній павутині, становлять інтерес для багатьох дисциплін, зокрема соціології, політології, психології, культурології, менеджменту, етнології, криміналістики, риторики та економіки, але не обмежуються ними. Із урахуванням цього комплексний аналіз віртуальної ідентифікації та самопрезентації в онлайн-спільнотах потребує міждисциплінарного підходу. Такий підхід необхідний для вивчення різних ідентичностей, технік самопрезентації та захисту вебособистості, які використовують учасники мережі у віртуальному середовищі.

2. Аналіз та постановка задачі

Видобування вебданих [12] є галуззю досліджень, яка швидко зростає та сконцентрована на вилученні цінної інформації та знань із вебджерел. За останні роки обсяг і складність вебданих різко зросли, що призвело до нагальної потреби в ефективніших і дієвіших методах видобування. Видобування вебданих – це збирання, аналіз та використання інформації із вебресурсів, який передбачає використання методів машинного навчання, статистичного аналізу, опрацювання природної мови та інших технологій для отримання знань із вебданих. Видобування вебданих корисне для багатьох цілей, зокрема пошуку інформації, прогнозування трендів, аналізу поведінки користувачів та покращення бізнес-стратегій. Воно також використовується у дослідженнях інформаційного контенту соціальних мереж, електронної комерції, маркетингу та в інших сферах. Сфера видобування вебданих стрімко розвивається, регулярно з'являються нові методи та додатки. Вебаналіз – це процес вилучення цінної інформації з вебджерел. Існує кілька видів вебаналізу, кожен із яких концентрується на різних аспектах вебданих. Одним з найважливіших напрямів досліджень у галузі вебаналізу є видобування вебконтенту [13], який передбачає вилучення цінної інформації з вебсторінок. Методи видобування вебконтенту використовують для вилучення структурованих даних із вебсторінок. Крім того, видобування вебконтенту застосовують для вилучення неструктурованих даних, таких як текст і зображення, які є корисними для аналізу настроїв, тематичного моделювання та інших застосувань. Ще однією важливою сферою досліджень у галузі вебаналізу є аналіз вебструктури, який зосереджується на аналізі структури гіперпосилань в інтернеті. Аналіз взаємозв'язків між вебсторінками, виявлення спільнот вебсторінок і виявлення аномалій у структурі гіперпосилань в інтернеті здійснюють методами аналізу вебструктури. Ще одним важливим напрямом досліджень у галузі вебаналізу є аналіз використання вебсторінок. Ця галузь зосереджена на аналізі даних про використання інтернету, таких як журнали вебсерверів, для виявлення закономірностей і тенденцій у поведінці вебкористувачів. За допомогою цього методу виявляють популярні вебсторінки, навігаційні шаблони користувачів і потенційні загрози безпеці. Існує кілька типів вебаналізу, кожен з яких може надати цінну інформацію про вебдані. Розуміння цих типів аналі-

тики та їх застосування може допомогти дослідникам і практикам приймати обґрунтовані рішення про те, які методи вибрати для виконання конкретного завдання.

Видобування даних соціальних мереж є новим напрямом досліджень, який зосереджений на вилученні цінної інформації з платформ соціальних мереж, таких як Facebook, Twitter та Instagram. Методи видобування інформації в соціальних мережах ефективні для аналізу контенту, який створили користувачі, виявлення популярних тем і вивчення поширення інформації в соціальних мережах.

Видобування вебданих важливе в різних сферах, таких як електронна комерція, маркетинг і охорона здоров'я. Наприклад, методи вебаналізу можна використовувати для вивчення поведінки споживачів в інтернеті, розроблення таргетованих рекламних кампаній [14] та виявлення потенційних ризиків для здоров'я, аналізуючи онлайн-контент. Видобування вебконтенту передбачає вилучення корисної інформації з вебсторінок, ураховуючи структуровані та неструктуровані дані. До структурованих даних належать імена, адреси та номери телефонів, а до неструктурованих – текст, зображення та відео. Методи інтелектуального аналізу контенту можна використовувати для вилучення цих даних із різними цілями, зокрема для аналізу настроїв, моделювання тем і рекомендацій щодо контенту. Аналіз вебструктури зосереджується на аналізі структури гіперпосилань в інтернеті [15]. Цей тип аналізу можна використовувати для виявлення закономірностей у тому, як вебсторінки посилаються одна на одну, виявлення спільнот вебсторінок і пошуку аномалій у структурі гіперпосилань. Методи аналізу вебструктури широко застосовують у пошуковій оптимізації та ранжуванні вебсторінок.

Аналіз використання вебресурсів передбачає аналіз журналів вебсервера для виявлення закономірностей у поведінці вебкористувачів. Методи видобування даних можна використовувати для виявлення популярних вебсторінок, вивчення навігаційних шаблонів користувачів і виявлення загроз безпеці. Цей тип аналізу зазвичай використовують в електронній комерції та дизайні вебсайтів.

Видобування даних із соціальних мереж [16] передбачає вилучення корисної інформації з платформ соціальних мереж, таких як Facebook, Twitter та Instagram. Методи видобування інформації в соціальних мережах можна використовувати для аналізу користувацького контенту, виявлення трендових тем і вивчення поширення інформації в соціальних мережах. Цей тип аналізу зазвичай застосовують у маркетингу та аналітиці соціальних мереж. Вебаналіз думок передбачає вилучення думок і настроїв, виражених у вебконтенті, наприклад, в оглядах, блогах і публікаціях у соціальних мережах. Методи видобування думок застосовні для виявлення позитивних і негативних настроїв, а також їх інтенсивності. Цей тип аналізу зазвичай використовується в маркетингових дослідженнях і розроблення продуктів.

Інтелектуальний аналіз вебконтенту [17] є сферою досліджень, яка зосереджена на вилученні цінної інформації з вебджерел. За останні роки обсяг і складність вебданих різко зросли, що призвело до нагальної потреби в ефективніших і дієвіших методах видобування. Одним з основних напрямів досліджень стосовно інтелектуального аналізу вебконтенту є вилучення структурованих даних, таких як імена, адреси та номери телефонів, із вебсторінок. Цей тип видобування поширений в інтеграції даних і вебпошуку, останніми роками він став об'єктом інтенсивних досліджень. Однак у поточних дослідженнях все ще є кілька прогалин, зокрема потреба в ефективніших і точніших методах вилучення даних, а також необхідність вирішення питань, пов'язаних із конфіденційністю та захистом даних. Ще однією сферою, на якій зосереджено увагу в дослідженнях з видобування вебконтенту, є видобування неструктурованих даних [18], таких як текст і зображення. Цей тип видобування зазвичай використовують в аналізі настроїв, моделюванні тем і рекомендаціях щодо контенту, серед інших застосувань. Попри значний прогрес у цій галузі, в поточних дослідженнях все ще існує кілька прогалин, зокрема потреба в точніших і ефективніших методах аналізу тексту і зображень, а також необхідність вирішення питань, пов'язаних з мовними та культурними відмінностями. Однією з переваг використання методів аналізу вебконтенту є те, що вони можуть допомогти організаціям і приватним особам приймати обґрунтовані рішення на основі інформації, отриманої із вебданих. Наприклад, аналізують вебконтент для визначення вподобань і поведінки

клієнтів, що потрібно організаціям для розроблення цілеспрямованіших маркетингових кампаній. Крім того, видобування вебконтенту можна використовувати для виявлення нових тенденцій і закономірностей у вебданих, що може допомогти окремим особам і організаціям випереджувати конкурентів у своїх галузях. Сфера інтелектуального аналізу вебконтенту швидко розвивається, регулярно з'являються нові методи та додатки. Цей огляд літератури висвітлює деякі з останніх тенденцій і розробок у цій галузі, а також прогалини в поточних дослідженнях і переваги використання методів видобування вебконтенту. Усуваючи ці прогалини та використовуючи переваги цього методу, дослідники та практики можуть продовжувати робити значний внесок у сферу видобування вебконтенту.

3. Інтелектуальний аналіз вебконтенту для формування соціально-цифрової ідентичності вебкористувача

Соціально-цифрова ідентичність охоплює соціально-демографічні характеристики вебкористувача, що проявляються в його онлайн-активності, поведінці та цифровому сліді. Цей термін визнає той факт, що присутність і взаємодія людини в інтернеті можуть розкривати важливі аспекти її особистої та соціальної ідентичності, які можуть бути неочевидними у взаємодії з фізичним світом. Поняття соціально-цифрової ідентичності охоплює різні соціально-демографічні характеристики, такі як вік, стать, місце проживання, освіта, професія, інтереси, вподобання та переконання, які можна вивести з цифрової взаємодії людини. Аналізуючи соціально-цифрову ідентичність, дослідники можуть отримати уявлення про те, як люди взаємодіють з онлайн-спільнотами та контентом і як на їхню поведінку в інтернеті впливають соціально-демографічні характеристики. Інтелектуальний аналіз вебконтенту широко використовують у різних сферах, таких як електронна комерція, освіта, охорона здоров'я та аналіз соціальних мереж. Однією з головних переваг інтелектуального аналізу вебконтенту є те, що він дає змогу організаціям отримувати цінну інформацію та знання з великих обсягів даних, доступних в інтернеті. Аналізуючи вебконтент, організації можуть краще зрозуміти потреби, вподобання та поведінку своїх клієнтів. Наприклад, компанії, що займаються електронною комерцією, можуть аналізувати відгуки клієнтів, щоб виявити тенденції та закономірності, які допоможуть їм покращити свої продукти та послуги. Аналогічно, навчальні заклади можуть аналізувати відгуки студентів, щоб поліпшити методи викладання та навчальні програми. Ще однією причиною популярності інтелектуального аналізу вебконтенту є його здатність допомагати організаціям приймати рішення на підставі даних. Аналізуючи вебконтент, організації можуть виявити закономірності та тенденції, скориставшись цим під час прийняття рішень. Наприклад, фінансові установи можуть використовувати аналіз вебконтенту для виявлення потенційного шахрайства або для оцінювання кредитоспроможності клієнтів. Медичні працівники можуть використовувати інтелектуальний аналіз вебконтенту для виявлення закономірностей у даних пацієнтів, які можуть допомогти їм покращити методи лікування.

Інтелектуальний аналіз вебконтенту можна використовувати для аналізу настроїв – процесу визначення настроїв або емоцій, що стоять за фрагментом тексту. Це може бути корисно для аналізу відгуків клієнтів, постів у соціальних мережах і новинних статей. Аналізуючи настрої вебконтенту, організації можуть краще зрозуміти, як їхні клієнти або широка громадськість ставляться до певного продукту, послуги чи події. Переваги інтелектуального аналізу вебконтенту численні та різноманітні, що робить його важливим напрямом досліджень і розробок у галузі науки про дані.

Формування компонентів для здійснення інтелектуального аналізу вебконтенту полягає у формуванні наборів індикаторів (мовних ознак соціально-демографічних характеристик мовлення вебкористувачів) для навчальної вибірки та передбачає здійснення певних кроків.

Здійснення автоматизованого пошуку маркерів. Для виявлення характеристик соціально-цифрової ідентичності учасників вебспільнот необхідно сформувати словник лінгвокомунікативних ознак інтернет-комунікації учасників вебспільнот маркерів, визначити методом дослідження фонетико-графічні, словотворчі й лексико-семантичні особливості мовлення учасників віртуального простору. Автоматизований пошук маркерів характеристик соціально-цифрової ідентичності

здійснюють за допомогою розробленого спеціалізованого програмного забезпечення. Проаналізуємо допис учасника вебфоруму “Дівочі посиденьки” – Gossip [19] на наявність лінгвістичних та графічних маркерів інтернет-комунікації. Множина маркерів визначає відповідність учасника вебфоруму певним значенням характеристик соціально-цифрової ідентичності.

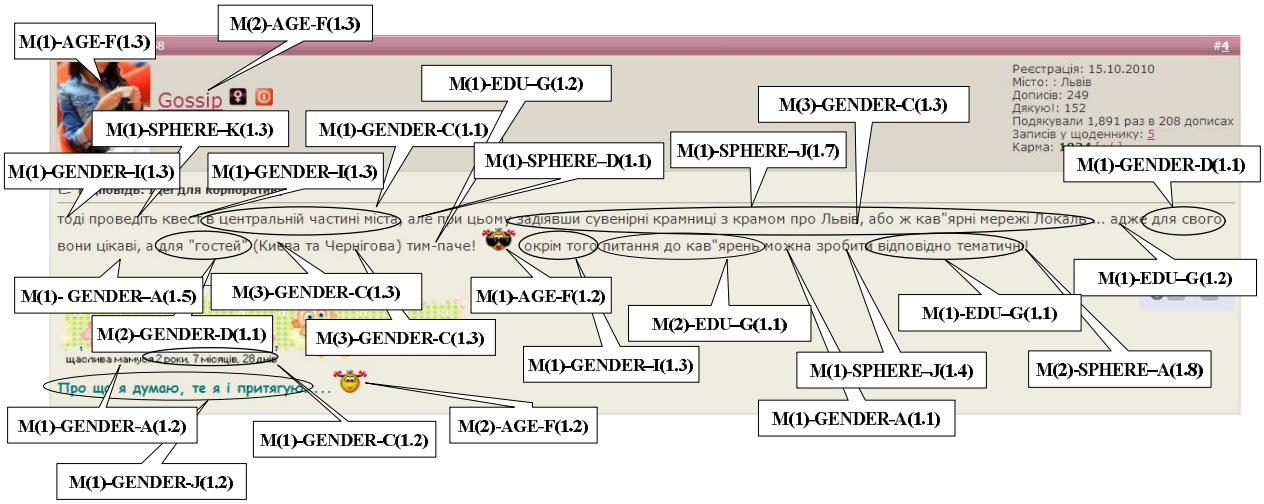

Рис. 1. Аналіз допису учасника вебфоруму “Дівочі посиденьки” – Gossip

У вищенаведеному дописі (див. рис. 1) методом аналізу виділено чотири групи маркерів. До *вікових маркерів* зарахуємо: маркери *M(1)-AGE-F(1.3)*, *M(1)-AGE-F(1.3)* індикативної ознаки *Аватари* та маркери *M(1)-AGE-F(1.2), M(2)-AGE-F(1.2)* індикативної ознаки *Графічні смайли*, які належать до лінгвокомунікативного індикатора ***Деформалізація***.

Маркери ґендеру: маркер *M(1)-GENDER-A(1.1)* індикативної ознаки *Модальні конструкції*, маркери *M(1)-GENDER-A(1.8)* та *M(2)-GENDER-A(1.8)* індикативної ознаки *Оклична інтонація, M(1)-GENDER–A(1.5)* індикативної ознаки *Аксіологічні модальні судження* та маркер *M(1)-GENDER-A(1.2)* індикативної ознаки *Згадка про емоції та почуття*, що належать до лінгвокомунікативного індикатора ***Емоційна складова***; *M(1)-GENDER-C(1.1)* індикативної ознаки *Просторові посилання* та маркери *M(1)-GENDER-C(1.3), M(2)-GENDER-C(1.3)* та *M(3)-GENDER-C(1.3)* індикативної ознаки *Географічні посилання*, маркер *M(1)*-GENDER-C*(1.2)* індикативної ознаки *Часові посилання*, які належать до лінгвокомунікативного індикатора ***Посилання***; *M(1)-GENDER-D(1.1)* та *M(2)-GENDER-D(1.1)* індикативної ознаки *Вказівка на особу, подію тощо*, які належать до лінгвокомунікативного індикатора ***Вказівка та інструкція***; *M(1)-GENDER–I(1.3)* та *M(2)-GENDER–I(1.3)* індикативної ознаки *Акцентування*, які належать до лінгвокомунікативного індикатора ***Сила, вплив, авторитетність***, *M(1)-GENDER-J(1.2)* індикативної ознаки *Пряме цитування*, яка належить до індикатора ***Збагачення мови***.

Маркери освіченості: маркер *M(1)*-EDU–G*(1.2)* індикативної ознаки *Пунктуаційні помилки*, маркери *M(1)*-EDU–G*(1.1)* та *M(2)*-EDU–G*(1.1)* індикативної ознаки *Синтаксичні помилки* та маркери *M(1)*-EDU–G*(1.3)* та *M(2)*-EDU–G*(1.3)* індикативної ознаки *Орфографічна помилка*, які належить до лінгвокомунікативного індикатора ***Мовні анормативи***.

Маркери сфери діяльності: маркер M(1)-SPHERE–K(1.3) індикативної ознаки *Наказовий спосіб*, яка належить до лінгвокомунікативного індикатора ***Воєнна сфера***; маркер M(1)-SPHERE–J(1.7) індикативної ознаки *Дієприслівникові й дієприкметникові звороти*, яка належить до лінгвокомунікативного індикатора ***Юридична сфера***; маркер M(1)-SPHERE–J(1.4) індикативної ознаки *Вживання дієслів теперішнього часу*, яка належить до лінгвокомунікативного індикатора ***Сільськогосподарська сфера***; маркер M(1)-SPHERE–D(1.1) індикативної ознаки *Складні речення з чітко вираженим складносурядним або складнопідрядним зв’язком*, які належать до лінгвокому-

нікативного індикатора ***Природнича сфера***. Схему процесу визначення маркерів віку характеристик соціально-цифрової ідентичності вебкористувача подано на рис. 2.

Рис. 2. Схема процесу визначення маркерів віку вебкористувача

Зважаючи на складність та великі витрати часу на аналіз інформаційних слідів учасників віртуальних спільнот, цей процес потребує автоматизації, що значно спрощує роботу та збільшує надійність отриманого результату.

Формування індикативних ознак. Індикативні ознаки формуються на основі маркерів зі спільними граматичними, лексико-семантичними та лексико-синтаксичними особливостями інтернет-комунікації учасників віртуальної спільноти. З метою диференції учасників віртуальної спільноти за значеннями характеристик експерти сформували множини вікових (рис. 3) та ґендерних (рис. 4) лінгвістичних ознак, ознак сфери діяльності (рис. 5) та освіченості (рис. 6) вебучасників на основі:

- досліджень, наукових теорій, ідеологій провідних вчених як вітчизняних, так і закордонних філологів, соціологів, лінгвістів, психологів, інформатиків;
- спеціалізованих словників (наприклад, комп'ютерно-мережевого жаргону, словників професійних термінів, словника молодіжного сленгу тощо);
- аналізу інформаційного наповнення україномовних віртуальних спільнот.

Рис. 3. Вікові лінгвокомунікативні індикатори

Відповідно до результатів дослідження, множину вікових лінгвокомунікативних індикаторів, яку формують вікові індикативні ознаки, подано на рис. 3.

Множину ґендерних лінгвокомунікативних індикаторів, яку формують ґендерні індикативні ознаки, наведено на рис. 4.

Рис. 4. Ґендерні лінгвокомунікативні індикатори

Множину лінгвокомунікативних індикаторів сфери діяльності, яку формують відповідні індикативні ознаки, наведено на рис. 5.

Фактично, лінгвістичні ознаки учасника віртуальної спільноти, які вказують на рівень освіти, це і є типові помилки в інтернет-комунікації.

Для пошуку та коригування цієї множини помилок науковці розробили багато алгоритмів як для англійської, так і для української мови, хоч і не настільки ґрунтовно. Зважаючи на це, розроблення нового автоматизованого засобу для пошуку і коригування помилок у вебконтенті не є вирішальним, оскільки оптимальним вирішенням цього завдання є відомий автоматизований засіб,

що ефективно функціонує [14], який полягає в аналізі тексту, фільтруванні слів, відбиранні слів з помилками та їх коригуванні.

Рис 5. Лінгвокомунікативні індикатори сфери діяльності

Множину лінгвокомунікативних індикаторів рівня освіченості та індикативні ознаки, які формують цю множину, подано на рис. 6.

Рис 6. Лінгвокомунікативні індикатори освіченості

Формування наборів лінгвокомунікативних індикаторів. Основне завдання цього процесу – консолідація лінгвістичних та комунікативних індикативних ознак інтернет-комунікації. Формування наборів лінгвокомунікативних індикаторів полягає у групуванні індикативних ознак в інтуїтивно-смислові групи. Візуалізацію результатів роботи подано в класифікації лінгвокомунікативних індикаторів для кожного значення характеристик (рис. 7).

Рис. 7. Схема класифікації лінгвокомунікативних індикаторів для кожного значення характеристики вебкористувача

Надалі у роботі будемо використовувати наведені у цих таблицях позначення для лінгвокомунікативних індикаторів та індикативних ознак.

Формування матриці лінгвокомунікативних індикаторів. На основі наборів лінгвокомунікативних індикаторів експерти формують матрицю лінгвокомунікативних індикаторів (1) методом комп'ютерно-лінгвістичного аналізу інформаційного наповнення віртуальних спільнот для кожного значення СД характеристики певної СД, що визначаємо окремо.

У результаті для кожного значення певних соціально-демографічних характеристик (СДХ) отримуємо матрицю лінгвокомунікативних індикаторів:

$$\boldsymbol{LingComInd}^{(\boldsymbol{SdCh,Vc})} = \begin{pmatrix} Ind_{1,1}^{(SdCh,Vc)} & \cdots & Ind_{1,j}^{(SdCh,Vc)} & \cdots & Ind_{1,N\_Vl(SdCh,Vc)}^{(SdCh,Vc)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Ind_{i,1}^{(SdCh,Vc)} & \cdots & Ind_{i,j}^{(SdCh,Vc)} & \cdots & Ind_{i,N\_Vl(SdCh,Vc)}^{(SdCh,Vc)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Ind_{N\_Ind(SdCh,Vc),1}^{(SdCh,Vc)} & \cdots & Ind_{N\_Ind(SdCh,Vc),j}^{(SdCh,Vc)} & \cdots & Ind_{N\_Ind(SdCh),N\_Vl(SdCh,Vc)}^{(SdCh,Vc)} \end{pmatrix}, \quad (1)$$

де N_V1 – функція, яка для кожної характеристики визначає кількість значень цієї характеристики соціально-цифрової ідентичності; N_Ind – функція, яка для кожного значення характеристики визначає кількість лінгвокомунікативних індикаторів цього значення характеристики соціально-цифрової ідентичності.

Кожен рядок матриці (2) є вектором лінгвокомунікативних індикаторів певної характеристики соціально-цифрової ідентичності:

$$Ind^{(SdCh,Vc)} = \left(Ind_{1,1}^{(SdCh,Vc)} \ldots Ind_{N\_Ind(SdCh,Vc),j}^{(SdCh,Vc)} \ldots Ind_{N\_Ind(SdCh),N\_V1(SdCh,Vc)}^{(SdCh,Vc)} \right). \quad (2)$$

Стовпець матриці (3) є вектором індикаторів певного значення характеристики соціально-цифрової ідентичності досліджуваної вебспільноти (див. таблицю):

$$LingComInd^{(SdCh,Vc)} = \begin{pmatrix} Ind_{1,1}^{(SdCh,Vc)} \\ Ind_{i,1}^{(SdCh,Vc)} \\ Ind_{N\_Ind(SdCh,Vc),1}^{(SdCh,Vc)} \end{pmatrix}. \quad (3)$$

Табличне подання функцій N_Vl та N_Ind

Соціально-демографічна характеристика *(SDCh)*	Кількість значень СДХ (*N_Vl*)	Кількість ЛК індикаторів (*N_Ind*)
Вік (Age)	2	6
Стать (Gend)	2	12
Рівень освіченості (Edu)	3	7
Сфера діяльності (Sphere)	3	11

За таким принципом формуємо матрицю для кожного вебучасника. Для обчислення відстані від еталонного значення характеристики соціально-цифрової ідентичності до кожного можливого значення характеристики соціально-цифрової ідентичності атомарного *k*-го учасника вебспільноти за основу беремо формулу визначення евклідової відстані (4):

$$\rho_j^{(k)}(Value,User) = \sqrt{\sum_{i=1}^{N\_Ind(SdCh,k)} \left(Ind_{i,j}^{(SdCh,Vc)} - Ind_{i,j}^{(SdCh,U)} \right)^2 * w_i^{(SdCh)}}, \quad (4)$$

де $k \in 1 \ldots N\_Vl(SdCh,Vc)$; $w_i^{(SdCh)}$ – ваговий коефіцієнт конкретного лінгвокомунікативного індикатора конкретного значення характеристики соціально-цифрової ідентичності.

Як результат вибираємо таке значення характеристики соціально-цифрової ідентичності, для якого справджується $\rho^* = \min(\rho_k)$. Причому матриця $LingComInd = (Ind_{ij})$ універсальна для всіх значень певної СДХ конкретної віртуальної спільноти, для якої синтезовано моделі. Залежно від тематики та виду віртуальної спільноти модель для кожного зі значень СДХ синтезують за допомогою автоматизованої інформаційно-аналітичної системи моніторингу.

Вагові коефіцієнти лінгвокомунікативних індикаторів (5) подано у векторі:

$$W^{(Vl,SdCh)} = \left(w_1^{(Vl,SdCh)} \quad \dots \quad w_j^{(Vl,SdCh)} \quad \dots \quad w_{N\_Ind(SdCh,Vc)}^{(Vl,SdCh)}\right) \tag{5}$$

Вектор вагових коефіцієнтів індикаторів характеристики соціально-цифрової ідентичності значення СДХ – Vl, отриманий як результат роботи автоматизованої інформаційно-аналітичної системи моніторингу.

Важливість лінгвокомунікативних індикаторів визначається ваговими коефіцієнтами. Результати аналізу різняться залежно від специфіки вебспільноти. Що більше значення коефіцієнта, то важливіший лінгвокомунікативний індикатор для верифікації відповідної характеристики соціально-цифрової ідентичності саме у конкретній віртуальній спільноті.

Визначення вагових коефіцієнтів лінгвокомунікативних індикаторів. Вагові коефіцієнти для лінгвокомунікативних індикаторів усіх значень характеристики соціально-цифрової ідентичності для кожної характеристики соціально-цифрової ідентичності визначають із використанням інформаційної системи багаторівневого комп'ютерного моніторингу [20]. На етапі формування масиву вхідних даних інформаційної системи багаторівневого моніторингу здійснюється опрацювання інформаційних слідів учасників вебспільнот на наявність у них маркерів з метою формування наборів лінгвокомунікативних індикаторів для конкретної віртуальної спільноти з відповідною тематикою.

Сформована матриця індикторів повинна відповідати низці вимог вхідних даних системи. Відповідно до вимог вхідних даних для інформаційної системи багаторівневого комп'ютерного моніторингу з метою синтезу якісної багатовимірної моделі масив вхідних даних повинен мати вигляд матриці числових характеристик об'єкта, інформативність якої повинна бути достатньою для синтезу якісних моделей [13]. Тобто ці вхідні дані мають бути у вигляді матриці частотних характеристик маркерів кожного лінгвокомунікативного індикатора в інформаційному сліді учасника віртуальної спільноти, що є основою для синтезу моделей в інформаційній системі багаторівневого комп'ютерного моніторингу [11, 13].

Загальновживаним та найпопулярнішим методом для опрацювання такого масиву даних є математико-статистичний метод опрацювання даних. Проте математико-статистичний метод не дає змоги реалізувати створення повноцінної інформаційної системи для верифікації персональних даних учасника віртуальної спільноти.

Формування соціально-цифрової ідентичності вебкористувача здійснено із використанням програмного засобу "Верифікатор соціально-демографічних характеристик вебучасника". Верифікатор соціально-демографічних характеристик учасника мережі [9] є програмним інструментом, який використовує набори лінгвістичних та комунікативних індикаторів інтернет-спілкування між членами вебспільноти для верифікації значень їхніх соціально-демографічних характеристик. Цей інструмент дає змогу забезпечити соціально-цифрову ідентичність учасника мережі. Для моделювання інформаційної схеми словника та програмного інструменту верифікації соціально-демографічних характеристик учасників вебспільноти використано засоби діаграмно-структурного моделювання. Сучасні вебтехнології застосовано під час розроблення інструментарію для формування соціально-цифрової ідентичності учасників віртуальних спільнот.

Верифікацію соціально-демографічних характеристик учасників вебспільноти, які є важливими для модерації спільноти та мають бути достовірними, здійснюють на основі формальної моделі характеристик СД. Компонент, що відповідає за цю функцію, потребує даних від двох компонентів: компонента формування наборів лінгвістичних та комунікативних індикаторів та компонента формування інформаційного сліду.

Компонент верифікації є логічним продовженням перевірки достовірності соціально-демографічних характеристик учасників віртуальної спільноти після завершення формування як мінімум базових лінгвістичних та комунікативних індикаторів та інформаційного сліду. Набори лінгвістичних і комунікативних індикаторів [11] формуються на основі постійно оновлюваного спеціалізованого словника маркерів.

Основні функції компонента перевірки соціально-демографічних характеристик – виявити неправдиву інформацію в акаунті учасника вебспільноти, виконавши комп'ютерно-лінгвістичний аналіз його інформаційного сліду. У разі недостатності даних для перевірки однієї або декількох соціально-демографічних характеристик учасника вебспільноти він генерує звіт для модератора віртуальної спільноти. Крім того, компонент повідомляє модераторові про неповноту спеціалізованого словника маркерів або про некоректність сформованих на його основі наборів лінгвістичних і комунікативних індикаторів.

4. Розроблення структури спеціалізованого словника для визначення соціально-демографічних характеристик вебкористувача

Соціально-демографічні характеристики (СДХ) визначаємо за допомогою маркерів, які містяться у спеціалізованому словнику та формують соціально-цифрову ідентичність вебкористувача. Структуру словника розроблено відповідно до схеми лінгвокомунікативних індикаторів СДХ.

Для моделювання інформаційної схеми розглянемо детально інформаційну схему словника, наведену на рис. 8, прокоментувавши сутності та їхні основні атрибути.

Верифікація великою мірою спирається на спеціалізований словник маркерів, який відіграє вирішальну роль у комп'ютерному лінгвістичному аналізі інформаційного сліду, який залишили члени вебспільноти. Без повної бази даних маркерів верифікація соціально-демографічних характеристик була б неможливою. Ці маркери необхідні для формування наборів лінгвістичних та комунікативних індикаторів, які вказують на соціально-демографічну характеристику учасника мережі.

Інформаційна модель окреслює вимоги до змісту словника маркерів. На основі цієї моделі, а також потреб і специфіки вебспільнот розроблено інтерфейс словника маркерів, необхідний для верифікації соціально-демографічних характеристик вебкористувачів.

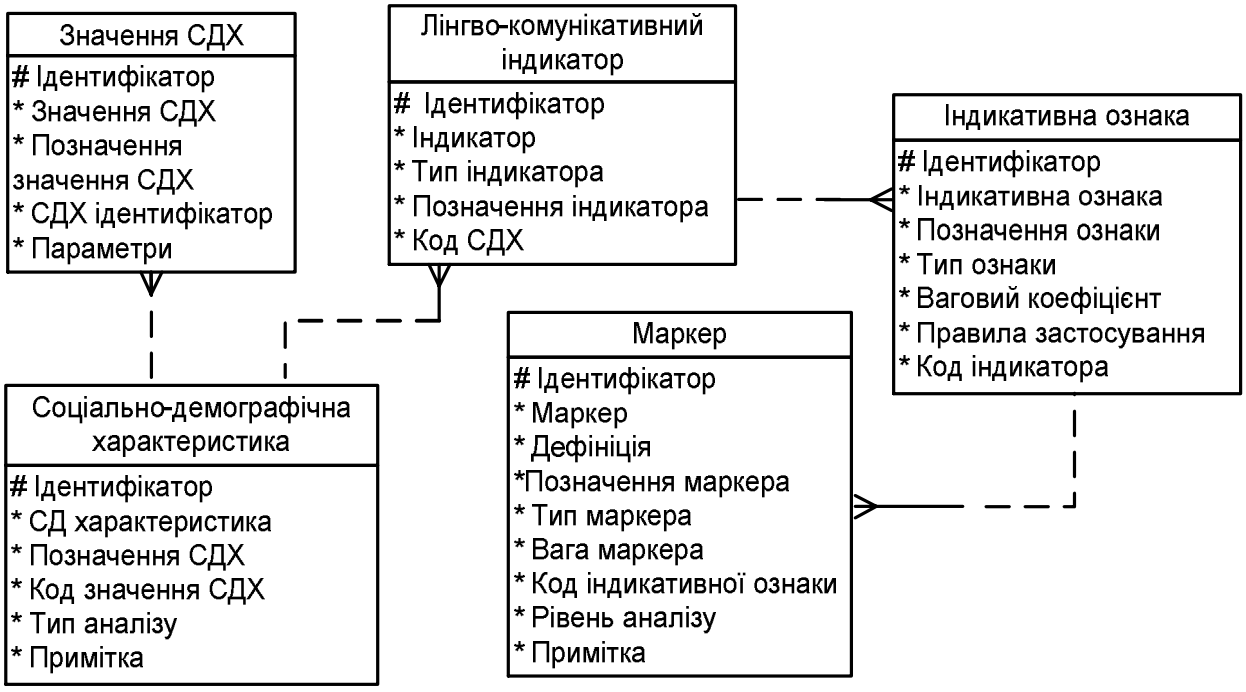

Рис. 8. Інформаційна схема словника маркерів характеристик вебкористувача

Спеціалізований словник соціально-демографічних характеристик членів віртуальних спільнот – це велика база даних маркерів, сформована за певним алгоритмом. Цей алгоритм передбачає уніфікацію, структурування та впорядкування великого масиву інформаційного контенту вебспільноти, використання методу комп'ютерного лінгвістичного аналізу та аналізу праць фахівців різних галузей для виявлення маркерів належності членів вебспільноти до певної соціально-демографічної характеристики, а також наповнення спеціалізованого інформаційного словника. Словникова база даних автоматично і систематично адаптується відповідно до тематики, специфіки віртуальної спільноти та інтернет-комунікації учасників, що входять до неї. Нарешті, словник інтегрований у програмний інструмент – верифікатор соціально-демографічних характеристик учасника-вебкористувача.

Користувацький інтерфейс спеціалізованого словника маркерів учасника вебспільноти подано на рис. 9.

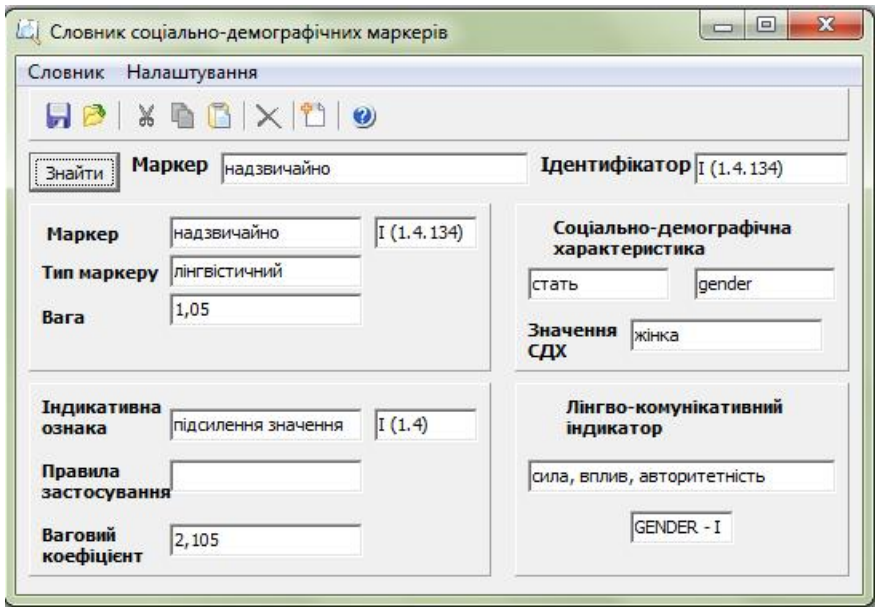

Рис. 9. Користувацький інтерфейс словника СД маркерів вебучасника

Розширена версія словника соціально-демографічних маркерів вебучасника за функціональністю та дизайном майже не відрізняється від звичайної версії, проте на інтерфейсі розміщено ще поле зі списком усіх наявних у базі соціально-демографічних маркерів та їх ідентифікаторів (див. рис. 10).

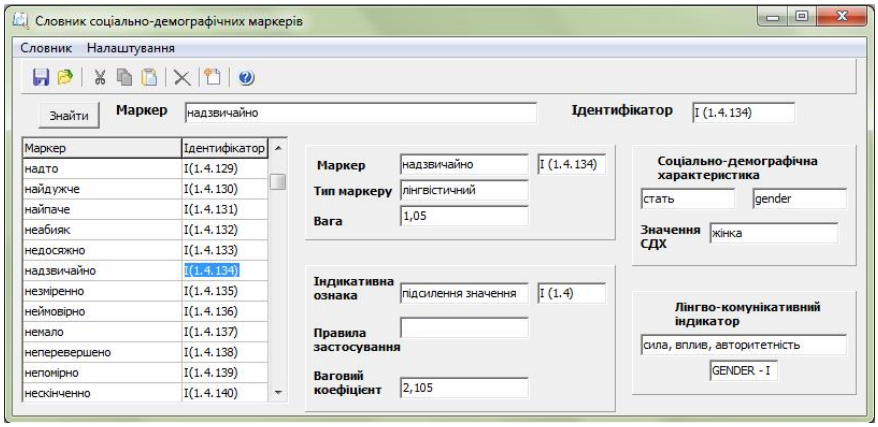

Рис. 10. Інтерфейс розширеної версії спеціалізованого словника маркерів контенту учасника вебспільноти

Підтримка наповнення спеціалізованого словника соціально-демографічних маркерів вебучасника в актуальному стані та дотримання чіткої структури лінгвокомунікативних індикаторів є ключовими чинниками якості та надійності результатів верифікації соціально-демографічних характеристик учасника віртуальної спільноти. Також від цих факторів залежать результати верифікації учасників віртуальної спільноти за допомогою програмного засобу "Верифікатор соціально-демографічних характеристик вебучасника".

Висновок

У сучасну цифрову епоху перевірка персональних даних вебкористувачів стала важливим завданням із різних причин, ураховуючи захист приватності, безпеку та дотримання нормативних вимог. Однак перевірка персональних даних – складне завдання, особливо у випадках, коли користувачі надають неправдиву або неточну інформацію. У цій статті запропоновано метод інтелектуального аналізу вебданих для перевірки персональних даних вебкористувачів на основі методу видобування вебконтенту для підвищення точності та ефективності верифікації. Цей метод передбачає залучення спеціалізованого словника маркерів контенту учасника вебспільноти для аналізу консолідованих вебданих, пов'язаних із персональними даними вебкористувача. За допомогою запропонованого методу зібрано дані профілів у соціальних мережах, записах у блогах та інші види загальнодоступної інформації, які перевіряють на наявність неправдивих або неточних персональних даних.

Переваги запропонованого методу подвійні. По-перше, він може істотно зменшити час і ресурси, необхідні для перевірки персональних даних, що робить його ефективнішим і економічно вигіднішим рішенням для організацій і приватних осіб. По-друге, він може забезпечити точніший і надійніший спосіб перевірки персональних даних, що може підвищити рівень захисту приватності та безпеки вебкористувачів.

Перевірка персональних даних вебкористувачів – важливе завдання, які має серйозні наслідки для захисту конфіденційності, безпеки та дотримання вимог законодавства. Запропонований метод використовує методи аналізу вебконтенту для підвищення точності та ефективності верифікації, забезпечуючи надійніший засіб перевірки персональних даних. Метод є цінним інструментом для організацій і приватних осіб, які потребують точної та надійної перевірки персональних вебданих.

Список використаних джерел

[1] K. Krippendorff, "Content analysis". SAGE Publications, Inc., 2019. DOI: 10.4135/9781071878781.

[2] S. C. Herring, "Web Content Analysis: Expanding the Paradigm. International Handbook of Internet Research". Springer, Dordrecht, 2009, pp. 233–249. DOI: 10.1007/978-1-4020-9789-8_14.

[3] P. Loyola, P. E. Roman, J. D. Velasquez, "Predicting web user behavior using learning-based ant colony optimization". Engineering Applications of Artificial Intelligence, 25(5), pp. 889–897, 2012.

[4] S. Fedushko, T. Ustyianovych, "E-Commerce Customers Behavior Research Using Cohort Analysis: A Case Study of COVID-19". Journal of Open Innovation: Technology, Market, and Complexity, 8(1), 12, 2022. DOI: 10.3390/joitmc8010012

[5] S. Fedushko, M. Davidekova, "Analytical service for processing behavioral, psychological and communicative features in the online communication". Procedia Computer Science, Vol. 160, pp. 509–514, 2019. DOI: 10.1016/j.procs.2019.11.056

[6] Y. C. Yang, "Web user behavioral profiling for user identification". Decision Support Systems, 49(3), pp. 261–271, 2010. DOI: 10.1016/j.dss.2010.03.001

[7] D. O'Neil, "Analysis of Internet users' level of online privacy concerns". Social Science Computer Review, 19(1), pp. 17–31, 2001. DOI: 10.1177/089443930101900103

[8] J. J. Hathaliya, S. Tanwar, "An exhaustive survey on security and privacy issues in Healthcare 4.0". Computer Communications, 153, pp. 311–335, 2020. DOI: 10.1016/j.comcom.2020.02.018

[9] С. С. Федушко, "Розроблення системи верифікації соціально-демографічних даних учасника віртуальної спільноти". Радіоелектроніка, інформатика, управління, № 3, С. 87–92, 2016. DOI: 10.15588/1607-3274-2016-3-11.

[10] M. L. Cantuaria, V. Blanes-Vidal, "Self-reported data in environmental health studies: mail vs. web-based surveys". BMC medical research methodology, 19(1), pp. 1–13, 2019. DOI: 10.1186/s12874-019-0882-x
[11] С. С. Федушко, Г. І. Білущак, "Формування системи лінгво-комунікативних індикаторів соціально-демографічних характеристик web-учасників". Управління розвитком складних систем: зб. наук. праць Київського нац. університету будівництва і архітектури, 18, С. 112–122, 2014. DOI: 10.6084/m9.figshare.11089865.v1
[12] S. Sharma, S. K. Sharma, "Web Mining: A Framework". IITM Journal of Management and IT, 12(1), pp. 93–98, 2021.
[13] Ibrahim, K. K., Obaid, A. J. "Web Mining Techniques and Technologies: A Landscape View". Journal of Physics: Conference Series. IOP Publishing, Vol. 1879, No. 3, p. 032125, 2021. DOI: 10.1088/1742-6596/1879/3/032125
[14] О. Я. Ярмолюк, О. С. Борисенко, Ю. В. Фісун, "Теоретико-методологічні аспекти таргетованої реклами як інструменту комплексного інтернет-маркетингу". Науковий вісник Херсонського державного університету. Серія: Економічні науки, (46), С. 23–29, 2022.
[15] Н. Гук, С. Диханов, І. Долотов, "Аналіз структури сайту з використанням поняття модулярності". Математичне та комп'ютерне моделювання. Серія: Фізико-математичні науки, С. 99–114, 2020.
[16] Y. Kryvenchuk, M. Y. Khanas, "Алгоритм видобування та опрацювання споріднених даних в соціальних мережах". Вісник Хмельницького національного університету, № 4, 2022 (311), С. 115–118, 2022. DOI: 10.31891/2307-5732-2022-311-4-115-118.
[17] О. І. Чумаченко, О. І. Житков, "Інтелектуальний аналіз веб-даних". Електроніка та системи управління, 2(32), С. 14–20, 2012.
[18] Ю. В. Рогушина, "Засоби та методи аналізу неструктурованих даних". Проблеми програмування, № 1, С. 57–77, 2019. DOI: 10.15407/pp2019.01.057.
[19] Веб-форум "Дівочі посиденьки". http://posydenky.lvivport.com/showthread.php?t=76784.
[20] С. В. Голуб, А. С. Авраменко, "Підвищення ефективності координації структури інформаційної системі комп'ютерного моніторингу з багаторівневим перетворенням даних". Секція 1. Сучасні аспекти математичного та імітаційного моделювання систем в екології, 21, С. 277–278, 2013.
[21] О. В. Харченко, С. В. Голуб, І. А. Жирякова, "Удосконалення методу висхідного синтезу елементів в інформаційній технології багаторівневого моніторингу мобільного робота". Математичні машини і системи, (3), С. 41–47, 2016.

INTELLECTUAL ANALYSIS OF WEB CONTENT FOR THE FORMATION OF SOCIAL AND DIGITAL IDENTITY OF WEB USER

Solomiia Fedushko

Lviv Polytechnik National University, 12, S. Bandery str., Lviv, 79013, Ukraine

The article discusses the growing need to analyze and understand web user behavior due to the unprecedented amount of digital content being generated and distributed on the Internet. Web content intelligence is presented as an effective way to explore and extract valuable information from online content, including websites, social media platforms, and other digital sources, to better understand web users' interests, preferences, and behaviors. The ability to identify web users based on their online behavior is important for industries such as marketing, psychology, and law enforcement. However, there are certain problems associated with this approach, such as ensuring the privacy and security of web users' data, as well as assessing the accuracy and reliability of web content analysis tools. The purpose of the article is to review the current state of web content analysis, its potential applications in various industries, and its role in shaping the digital future. The article emphasizes the importance of an interdisciplinary approach to the study of virtual identification and self-presentation in online communities, taking into account the socio-demographic characteristics of a web personality involved in social interactions. The article also explores the latest trends and developments in the field of web data mining, including web content analysis, web structure analysis, web page usage analysis, and social media data analysis. A software solution for conducting intelligent analysis of web content is proposed to form a social and digital identity of a web user using a specialized dictionary of content markers of a web community member.

Key words: *web content; web data mining; social networks; Internet; intelligent analysis; social and digital identity.*