© Syerov Yu. Tracking of dangerous content of online communities as a key aspect of information-psychological security of online users // Ukrainian Scientific Journal of Information Security, 2017, vol. 23, issue 2, p. 113-121.

КОНКУРЕНТНА РОЗВІДКА ТА УПРАВЛІННЯ ЗНАННЯМИ / BUSINESS INTELLIGENCE & KNOWLEDGE MANAGEMENT

DOI: 10.18372/2225-5036.23.11764

ВІДСТЕЖЕННЯ ПОЯВИ НЕБЕЗПЕЧНОГО КОНТЕНТУ ОНЛАЙН СПІЛЬНОТ ЯК КЛЮЧОВИЙ АСПЕКТ ІНФОРМАЦІЙНО-ПСИХОЛОГІЧНОЇ БЕЗПЕКИ ОНЛАЙН-КОРИСТУВАЧІВ

Юрій Сєров

Національний університет «Львівська політехніка», Україна

СЄРОВ Юрій Олегович, к.т.н.

Рік та місце народження: 1981 рік, м. Сокаль, Україна.
Освіта: кандидат технічних наук за спеціальністю «Математичне та програмне забезпечення обчислювальних машин та систем».
Посада: доцент, заступник завідувача з наукової роботи кафедри соціальних комунікацій та інформаційної діяльності.
Наукові інтереси: методи протидії інформаційним війнам, методи та засоби побудови віртуальних спільнот та їхнє управління, проблеми побудови та розвитку інформаційного суспільства та електронне урядування.
Публікації: понад 100 наукових та навчально-методичних публікацій, серед них монографія, навчальні посібники, наукові статті у закордонних та вітчизняних авторитетних виданнях, матеріали конференцій.
E-mail: Yurii.O.Sierov@lpnu.ua

Анотація. *У статті запропоноване вирішення актуальної задачі розроблення методів відстеження появи небезпечного контенту для користувачів онлайн спільнот з метою підвищення їхньої інформаційно-психологічної безпеки. Впровадження отриманих результати в роботу онлайн спільноти дозволяють уникнути випадків притягнення до адміністративної та кримінальної відповідальності власників, адміністраторів, модераторів та пересічних користувачів онлайн спільнот. Методи недопущення, виявлення та розв'язання конфліктів між учасниками онлайн спільноти, метод відстеження появи небажаного інформаційного наповнення та метод застосування санкцій до учасників онлайн спільноти розроблено з метою відстеження появи небезпечного контенту для користувачів онлайн спільнот. Запропоновані методи суттєво підвищать інформаційно-психологічну безпеку користувачам, адміністраторам та модераторам онлайн спільнот з жорсткою структурою, строгою ієрархією та своєрідною спеціалізацією, а саме закритих корпоративних онлайн спільнот, онлайн спільнот навчальних курсів, вищих навчальних закладів та шкіл, державних, військових та поліцейських установ.*

Ключові слова: *онлайн спільнота, небезпечний контент, користувач, відстеження загроз, інформаційне наповнення, фільтрація забороненого контенту.*

Вступ

В час активного розвитку онлайн комунікацій у світовій юриспруденції зафіксовані прецеденти, пов'язані з протиправною діяльністю користувачів онлайн спільнот. Зокрема, такі випадки стосувалися притягнення до адміністративної та кримінальної відповідальності власників, адміністраторів, модераторів та пересічних користувачів онлайн спільнот з різних причин: розпалювання міжетнічної ворожнечі, особисті образи, тощо. Однак, часто звинувачення стосуються рядових учасників онлайн спільноти, які своїми діями спровокували конфлікти, то існують також прецеденти, коли відповідальність за інформаційне наповнення онлайн спільноти покладалась на адміністраторів та модераторів онлайн спільноти. Наприклад, Гамбурзький суд виніс обвинувачуваний вирок власнику онлайн спільноти [1], оскільки суддя прирівняв факт не видалення інформації на форумі з підтриманням модераторами та адміністр-

раторами позиції учасника. Законодавчі органи різних країн пропонують також притягувати до кримінальної відповідальності адміністраторів сайтів новин за повідомлення, які залишають відвідувачі.

Усі наведені факти свідчать про те, що у наш час власники, адміністратори та модератори повинні бути уважними і ретельно стежити за відповідністю інформаційного наповнення їх онлайн спільнот чинному законодавству, оскільки через злочинні дії одного з учасників спільноти можуть постраждати інші учасники, модератори, адміністратори та власники онлайн спільноти, а сам веб-ресурс може припинити своє існування. Наприклад [2, 3], у 2009 році на сайті «Юстініан» на форумі (де відвідувачі можуть звертатися по допомогу при вирішенні юридичних питань) користувач розмістив статтю: у ній низку осіб звинувачено в злочинній і рейдерській діяльності. Суди визнали винними власників інтернет-ресурсів (організацію «Юстініан» і «Українську правду») за анонімні дописи інших людей у зонах вільного доступу за поширення інформації, яка ганьбить честь, гідність і ділову репутацію позивача.

Наприкінці 2013 року Європейський суд з прав людини у справі Делфі АС проти Естонії, визнав правомірність застосування до власника вебсайту заходів цивільно-правової відповідальності за образливі коментарі, залишені користувачами на сайті новин.

Окрім згаданих випадків, адміністраторам та модераторам передусім необхідно остерігатися появи у онлайн спільноті інформаційного наповнення [4-6], що порушує авторські права (програмне забезпечення, відео- та аудіо продукція, книг), порнографічних матеріалів, шкідливого програмного забезпечення («вірусів»), інформації, що може спричинити конфлікти на міжрасовому, міжетнічному та міжрелігійному ґрунті. Таке інформаційне наповнення називатимемо небажаним для спільноти, або просто небажаним.

Дотримання усіх законів сприятиме стабільній роботі онлайн спільноти та комфортній атмосфері спілкування особливо важливим є для таких чітко структурованих онлайн спільнот [7-8]:

– закритих корпоративних онлайн спільнот;

– онлайн спільнот навчальних курсів, вищих навчальних закладів та шкіл;

– онлайн спільнот з державних, військових та поліцейських установ тощо.

Наведені вимоги ставлять адміністраторів та модераторів у складне становище. Навіть за невеликих обсягів приросту інформаційного наповнення робота з відстеження та виявлення небажаного інформаційного наповнення [10-12] потребує великої кількості часу і зусиль. За більших обсягів приросту інформаційного наповнення розв'язання цієї задачі без автоматизованих засобів стає нереальним.

Зважаючи на все вище сказане, метою цієї роботи є аналіз актуальних досліджень та розроблення ефективних методів фільтрації за чітко визначеними категоріями соціально-небезпечного інформаційного наповнення, яке генерують учасники онлайн спільнот.

Аналіз останніх досліджень і публікацій

Найбільш суворі регулювання діяльності інтернет-сервісів запроваджено у Китаї – це є розроблений фахівцями концепт «Великий китайський файрвол». У 2003 р. Ден Сяопін розробив «проект Золотий Щит» – систему цензури і фільтрації сервісів Інтернету. Проект є системою серверів на інтернет-каналі між провайдерами і міжнародними мережами передачі інформації, що фільтрує інформацію. За даними Г. Уолтона [13] – фахівця з Міжнародного центру з прав людини та демократичного розвитку, Китай реалізував найскладнішу фільтрацію інтернет-контенту, що здатна ефективно фільтрувати контент з використанням безлічі методів регулювання і технічного контролю (блокування IP-адрес і фільтрація контенту; фільтрація DNS і URL; скидання з'єднання тощо).

Відповідно до міжнародної інтернет-політики використання автоматизованих систем, які призначені для захисту користувачів Інтернету від інформаційного наповнення, що порушує авторські права, порнографічних матеріалів, шкідливого програмного забезпечення, інформації, що може спричинити конфлікти на міжрасовому, міжетнічному та міжрелігійному ґрунті, є одним з ефективним засобів боротьби з кібер-криміналом [14]. На сьогодні існує два методи боротьби з появою небезпечного контенту онлайн спільнот, такі як:

– блокування веб-сайтів та інформаційних сервісів у мережі Інтернет;

– фільтрація інформаційного наповнення.

Блокування веб-сайтів та інформаційних сервісів у мережі Інтернет полягає у свідомому блокуванні усіх інтернет-ресурсів, контент яких є небезпечним та заборонений правилами адміністрації цих ресурсів. Аналіз контенту веб-сайтів здійснюють спеціалізовані програми обмеження вебконтенту – контент-фільтр.

Результатом цього методу є обмеження користувачу в цілковитому доступі до певних сайтів або послуг мережі Інтернет.

Контент-фільтр працює по статистичному принципу, тобто підраховує заздалегідь певні слова тексту і визначає категорію, до якої належить вміст сайту. Метою таких програм є обмеження доступу в Інтернет для шкіл, підприємств, релігійних організацій і т.д. Найчастіше контент-фільтри використовуються для обмеження доступу для дітей і підлітків, в навчальних закладах, бібліотеках і на робочих місцях в різних установах, а також ігрових клубах та інтернет-кафе.

Функція контент-фільтра аналізує текст інтернет-сторінки і якщо на ній перевищено поріг присутності заборонених слів в тексті, то перегляд такої сторінки автоматично блокується. Великий список заборонених слів для контент-фільтрації вже занесений в програму для батьківського контролю ChildWebGuardian PRO. Вони розбиті на категорії (порнографія, насильство, наркотики, тероризм і екстремізм та ін.), але користувач може доповнювати ці списки своїми словами, налаштовуючи правила контент-фільтрації під свої потреби.

Для блокування порнографії та інших класифікацій матеріалів з окремих комп'ютерів або мереж доступно різноманітне програмне забезпечення для контролю контенту, батьківського контролю та фільтрації. Прикладами комерційних веб-фільтрів [15-17] є Content Protect, Cyber Patrol, Cyber Sentinel, Filter Pak, McAfee Parental Controls, Cyber Sitter, Norton Parental Controls, Bess, Net Nanny, SeeNoEvil, SurfWatch та інші. Peacefire є однією з найбільш помітних клірингових центрів для таких контрзаходів.

Розглянемо ще один більш трудоємкий, але безкоштовний клієнтський спосіб фільтрації Інтернет контенту [18, 19]. Найрозповсюдженішим інтернет-фільтром є інтернет-цензор. В основі роботи програми лежить технологія «білих списків». Основною функцією інтернет-фільтру є блокування доступу до інтернет-сайтів, які не входять до дозволеної бази сайтів («білий список») або блокування сайтів, що знаходяться в базі заборонених сайтів («чорний список»). Списки сайтів заповнюються вручну, хоча є можливість завантажити списки, створені компанією розробником.

Попри зазначені переваги застосування методу блокування веб-сайтів та сервісів у мережі Інтернет, необхідно звернути увагу на недоліки. Важливими недоліками застосування зазначеного методу є:

– не здійснюють фільтрацію та моніторинг чатів, відкритих бесід та повідомлень;

– більшість програмних засобів одразу блокує користувача або інтернет-ресурси;

– немає можливості детального налаштування;

– іноді помилка в оцінці небезпеки сайту призводить до обмеження доступу до нешкідливою інформації.

Фільтрація інформаційного наповнення

Популярні соціальні медіа-сайти, такі як Facebook [20] та Twitter [21, 22], цензурують публікації, що містять порушення, такі як:

– мова ворожнечі або пропаганда проти особи, організації або групи осіб на основі расової належності, національності, етнічного походження, кольору шкіри, віросповідання, інвалідності, віку, статі, сексуальної орієнтації, гендерної ідентичності, статусу ветерана чи іншого захищеного статусу;

– насильство або загроза насильства над людьми чи тваринами;

– прославляння шкоди або пов'язаного з ним інформаційного контенту;

– організації або особи, пов'язані з пропагандою ненависті, злочину або пов'язаного з тероризмом інформаційного наповнення;

– агресивний контент, який може спричинити сильну негативну реакцію або завдати шкоди;

– нав'язливе, вульгарне, образливе або нецензурне інформаційне наповнення.

З 2016 року Facebook і Twitter [23] додержуються політики фільтрації інформаційного наповнення, яке містить вищенаведені ознаки протягом 24 годин.

Специфікою цього методу є автоматичне сканують повідомлень або коментарів користувачів, а також автоматична заміна чи цензура окремих слів або фраз за допомогою розроблених скріптів – словофільтрів (wordfilter). Ці засоби зазвичай використовують для модерування інформаційного наповнення інтернет-форумів та чатів.

Більшість розроблених словофільтрів шукають тільки певні літери у рядку, а також видаляють або перезаписують їх незалежно від контексту. Вдосконалені словофільтри роблять деякі винятки для контексту, а найдосконаліші словофільтри використовують регулярні вирази [24].

Прикладом використання словофільтру у онлайн спільноті «File SharingTalk» [25] є заміна літер забороненого слова на символи «*» (рис. 1).

Рис. 1. Приклад використання словофільтру у онлайн спільноті File SharingTalk

Основна проблема використання слово фільтрів є їхній вплив на слова, які не призначені для фільтрування. Випадково можуть стати причиною псування смислу інформаційного наповнення. Типовою проблемою застосування цього методу є фільтрування коротких словів. Наприклад [26]: «Do you need istance for playing clical music?». Кілька слів можуть бути відфільтровані, якщо пробіл проігноровано, внаслідок чого «as suspected» змінено на «uspected». Заборона вимовляти таку фразу, як «hard on», призведе до фільтрації нешкідливих повідомлень, таких як «That was a hard one!». І результатом фільтрації «Sorry I was hard on you» є такі фрази «That was a e!» та «Sorry I was you».

Як бачимо, цей метод є недосконалим і переважно адаптований для інформаційного наповнення написаного англійською мовою.

Розроблення методу відстеження появи небажаного інформаційного наповнення

Для зменшення ресурсоємності процесу модерування онлайн спільноти розроблено метод відстеження появи небажаного (для власників) інформаційного наповнення.

Метод відстеження появи небажаного інформаційного наповнення передбачає (рис. 2):

– фільтрацію текстового інформаційного наповнення;

– відстеження та класифікацію зовнішніх посилань;

– відстеження прикріплених файлів.

Фільтрація небажаного текстового інформаційного наповнення є найпростішим завданням із переліченого і полягає у знаходженні у новому інформаційному наповненні заборонених слів, видаленні їх чи заміні на інші слова, символи тощо.

База заборонених слів формується на початку створення онлайн спільноти і доповнюється упро-

довж усього часу її існування. Її створюють адміністратори залежно від вибраного сценарію розвитку онлайн спільноти.

Прикладом заборонених у онлайн спільнотах слів є ненормативна лексика, вислови, що ображають людей на расовому, етнічному, релігійному ґрунті тощо.

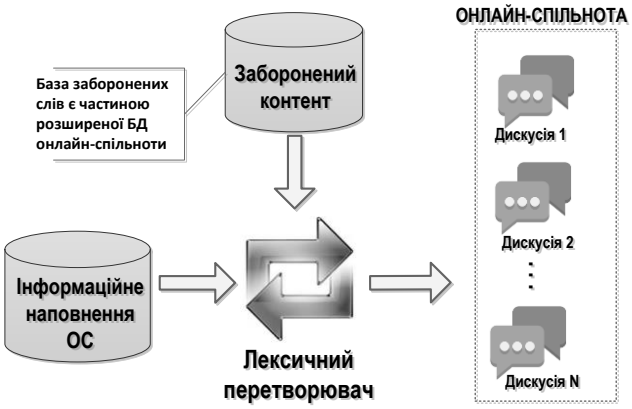

Рис. 2. Схема фільтрації забороненого контенту

Ще однією важливою складовою відстеження небажаного інформаційного наповнення є відстеження зовнішніх посилань з їх подальшою класифікацією та фільтрацією.

Відстежувати небажані зовнішні гіперпосилання складніше ніж текстове інформаційне наповнення, оскільки небажаних сайтів значно більше, ніж небажаних слів.

Перевірка інформаційного наповнення, на яке вказує гіперпосилання, також потребує участі людини. Адміністратор онлайн спільноти класифікує за допомогою автоматизованих засобів веб-сайти та сторінки, на які вказують гіперпосилання, зараховуючи їх до «чорного» (забороненого) чи «білого» (дозволеного) списку.

Гіперпосилання, що містяться у забороненому списку, фільтруються і не відображаються у дискусіях онлайн спільноти.

Відстеження, класифікація та фільтрація зовнішніх посилань відбувається за схемою, зображеною на рис. 3.

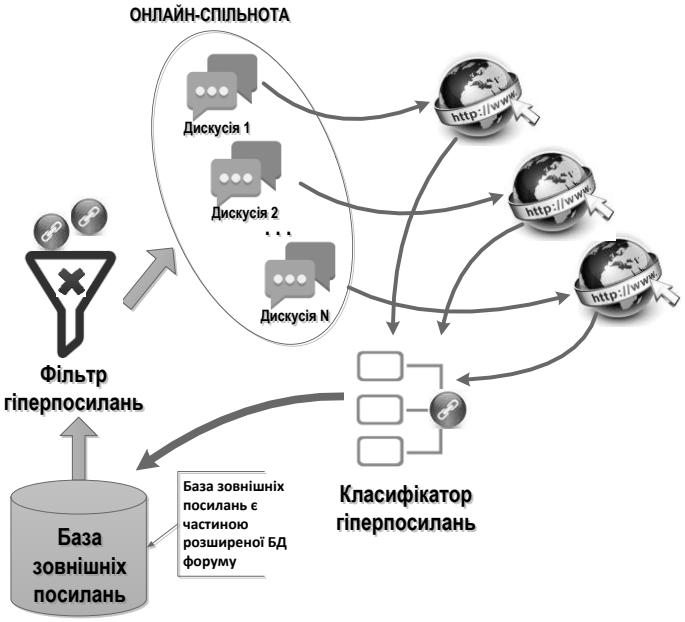

Рис. 3. Схема класифікації і фільтрації зовнішніх гіперпосилань

Алгоритм дії методу реалізується з урахуванням таких положень:

– небажане інформаційне наповнення може міститися у прикріплених файлах та гіперпосиланнях на зовнішні інформаційні ресурси, тому у разі появи нових повідомлень, що містять прикріплені файли чи зовнішні гіперпосилання, автоматизований засіб повідомлятиме адміністратора та модератора відповідного розділу про необхідність перевірки дискусії, у якій з'явилося повідомлення з новим інформаційним наповненням;

– зберігання даних про джерела та властивості небажаного інформаційного наповнення (назви файлів, заборонені словосполучення у назвах, заборонені зовнішні інформаційні ресурси) – «чорний» список – та дозволеного інформаційного наповнення – «білий» список;

– перевірка текстового інформаційного наповнення засобами фільтрування та лексичного аналізу одразу після його створення;

– підготовка необхідних даних для застосування санкцій до учасників, що створили небажане інформаційне наповнення, відповідно до встановлених у онлайн спільноті правил.

На основі цього алгоритму автором розроблена утиліта «Веб-цензор» (рис. 4). Розглянемо, як вона реалізується:

1. Після створення учасником онлайн спільноти нового інформаційного наповнення до нього застосовується фільтр забороненої лексики.

2. У разі виявлення заборонених у спільноті слів (наприклад, нецензурної лексики, образливих назв тощо) під час відображення у дискусіях онлайн спільноти їх заміняють на задані синоніми чи узагалі не відображають. Інформація про заборонені слова фіксується у звіті модератору.

3. Перевірка наявності зовнішніх посилань.

4. У разі виявлення зовнішніх гіперпосилань відбувається перевірка їх наявності у «чорному» списку.

5. У разі виявлення гіперпосилань у «чорному» списку гіперпосилання видаляються, інформація про порушення фіксується у звіті модератору. Якщо гіперпосилання ще не класифіковані, інформація про необхідність їх класифікації вноситься у звіт модератору.

6. Перевірка наявності прикріплених файлів.

7. Якщо виявлені прикріплені файли, інформація про це фіксується у звіті.

8. Звіт про виявлені порушення, некласифіковані зовнішні гіперпосилання та прикріплені файли надсилається адміністраторам.

На основі алгоритму створено автоматизований засіб виявлення небажаного та підозрілого інформаційного наповнення, результатом роботи якого є звіт адміністратору із завданнями, виконання яких потребує його участі – класифікації гіперпосилань та прикріплених файлів.

Вихідні дані алгоритму, зокрема інформація про виявлені порушення є вхідними даними для алгоритму застосування санкцій до учасників.

Оскільки спілкування у середовищі Веб 2.0 [26, 27] та, зокрема, на форумах ведеться у письмовому вигляді, доцільно вживати заходів для запобігання конфліктам на основі модерації [28-31] контенту повідомлень користувачів. Найпоширенішим видом конфлікту на форумах є «флейм» – словесна війна, що виникає у ході спілкування і дуже часто не пов'язана з першопричиною суперечки. Повідомлення можуть містити особисті образи і бути націленими на подальше розпалювання сварки.

Відсутність безпосереднього контакту з віртуальними співрозмовниками дає змогу користувачам вільніше висловлювати свою думку та відстоювати позицію, що частіше ніж в реальному спілкуванні, стає основою для виникнення конфліктів. Відзначається використання категоричних тверджень, демонстрація неповаги до аргументів співрозмовників та до них самих, неготовність визнати свої помилки, негативне оцінювання аргументів співрозмовників тощо.

Некоректно написане інформаційне наповнення часто є причиною породження конфліктів між учасниками онлайн спільноти. З метою недопущення, виявлення та розв'язання конфліктів між учасниками онлайн спільноти конфлікти у інтернет спільнотах класифіковано на такі типи:

Передконфліктна ситуація. Передконфліктна ситуація являє собою створення нової теми або появу допису, який може чимось зачепити інших учасників спільноти, що будуть ознайомлюватися з дописом чи дискусією. Найчастіше в передконфліктних повідомленнях та дописах трапляються невдалі і двозначні жарти, натяки тощо, образливі за нормативного тлумачення в серйозному сенсі, різкі висловлювання на адресу «сторонніх об'єктів» (ігор, корпорацій, спортивних команд, програм, фільмів, артистів), неаргументована критика, полемічні прийоми підміни понять, вияви «зверхнього ставлення», спричинені помилковим визначенням віку або кваліфікації співрозмовника, різні погляди на політичні та історичні події.

Відкритий конфлікт. Початок конфлікту. Якщо на повідомлення, почали реагувати інші учасники, створюючи свої повідомлення, які містять подібне інформаційне наповнення (наведений вище), вважають, що почався конфлікт інтересів, думок та поглядів.

1. Розвиток конфлікту. Ця стадія не обов'язково наявна у всіх конфліктах, що виникають під час спілкування. Залежно від швидкості реагування адміністраторів та модераторів онлайн спільноти на небажані повідомлення або від наміру учасників (чи учасника) продовжувати суперечку, конфлікт, може відразу перейти до завершальної стадії. Проте якщо ніяких заходів з боку адміністрації онлайн спільноти не вжито і учасники продовжують «словесні битви», до яких долучаються інші учасники, що поглиблюють проблему, то конфлікт розвивається.

2. Завершення конфлікту. Завершення конфлікту має кілька підвидів:

– вичерпання самої причини конфлікту, якщо учасники дійшли згоди і порозумілися щодо конфліктного питання;

– втручання адміністрації онлайн спільноти для розв'язання конфлікту;

– небажання учасників продовжувати суперечку.

Післяконфліктний період. Оскільки в онлайн спільнотах користувачів є поняттям нечітким, післяконфліктний період не так сильно виражений, як у реальному спілкуванні. Його тривалість може бути як дуже незначною, так і доволі великою. Щоб визначити тривалість післяконфліктного періоду, здійснюють модерацію повідомлень користувачів, між якими виник конфлікт, в інших дискусіях. Якщо напруження зберігається і в інших дописах, не пов'язаних з предметом конфлікту, слід очікувати виникнення нових конфліктів за інерцією.

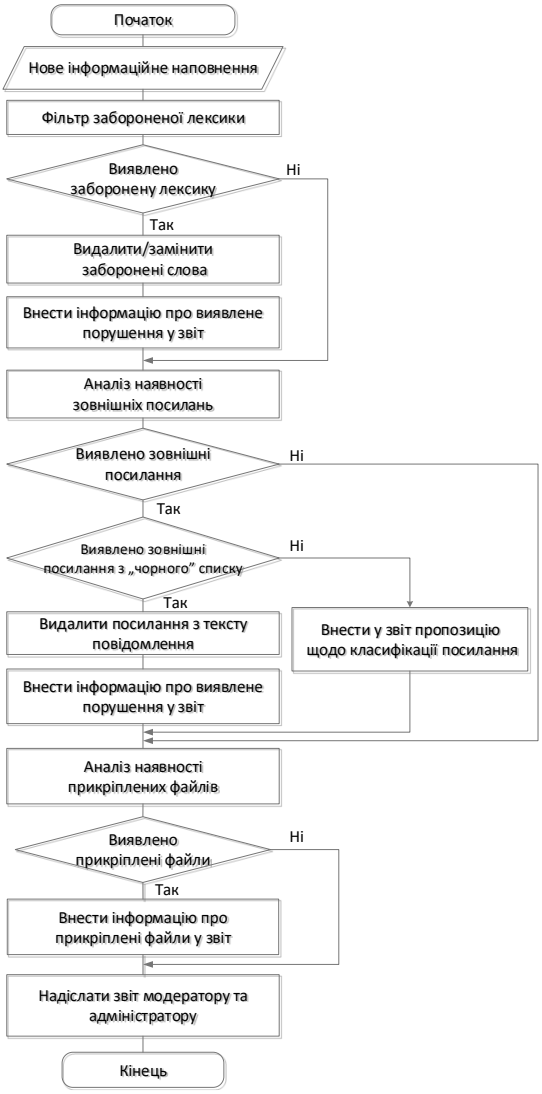

Рис. 4. Блок-схема утиліти «Веб-цензор»

Виділяють такі види конфліктів:

1. Міжособистісні конфлікти, тобто конфлікти між користувачами веб-форуму як між окремо взятими елементами.

2. Між особистістю і групою. Виникають здебільшого через так званий «тролінг», тобто намагання однієї людини завадити роботі онлайн спільноти, розміщуючи провокаційні повідомлення, щоб спричинити сварку та створити конфліктну ситуацію.

3. Міжгрупові. У онлайн спільнотах, як і в реальному житті, утворюються умовні групи «однодумців». Міжгрупові конфлікти виникають між двома або більше групами осіб, кожна з яких відстоює свої погляди на те чи інше питання. Керувати міжгруповим конфліктом надзвичайно важко, оскільки в них задіяна велика кількість людей.

Під час конфлікту його сторони можуть висловити нові ідеї, запропонувати цікаві способи вирішення тієї чи іншої проблеми. Проте конфлікти, що виникають часто та вчасно не знешкоджуються під час спілкування, можуть призвести до руйнування спільноти онлайн спільноти, зменшення кількості дописів, втрати популярності, відпливу учасників тощо.

Основні способи вирішення конфлікту:

– видалення учасника/учасників конфлікту;

– усунення об'єкта конфлікту;

– зміна позицій сторін конфлікту;

– доручення до конфлікту нового учасника, здатного завершити його за допомогою примусу;

– втручання у конфлікт модераторів та адміністраторів.

Для модератора онлайн спільноти важливою є не тільки модерація повідомлень користувачів, але й можливість визначити «тип учасника» для того, щоб мати змогу передбачити його поведінку в конфліктній ситуації ще до того, як така ситуація виникне.

Пропонуємо рекомендації модераторам та адміністраторам для уникнення конфліктних ситуацій та усунення наслідків конфліктів, що виникають в ході спілкування у онлайн спільноти середовища Веб 2.0:

1. Сформулювати правила поведінки учасників спільноти з урахуванням якомога більшої кількості можливих порушень та конфліктних ситуацій.

2. Попередити учасників про можливі санкції за порушення того чи іншого правила поведінки.

3. Формувати активну позицію учасників спільноти: заохочувати спільноту звертатися до адміністрації у разі виникнення конфліктної ситуації чи провокацій з боку певних учасників спільноти.

4. Стежити за дотриманням учасниками правил спільноти.

5. Окрім запропонованої схеми класифікації, створити систему поміток, за допомогою якої фіксувати «попередження» учасників і відстежувати дії «проблемних» учасників спільноти.

6. Приділяти особливу увагу, а за необхідності не допускати виникнення дискусій, які не стосуються тематики онлайн спільноти, можуть спровокувати конфлікт на грунті расових, релігійних, етнічних, гендерних відмінностей.

7. Виявивши сварку (флейм), одразу припиняти її.

8. Закривати дискусії, у яких виник конфлікт, або видаляти частину дискусії, яка не стосується її тематики.

9. Вчасно застосовувати необхідні адміністративні санкції: попередження, блокування доступу до форуму на певний час, безстрокове блокування доступу. Санкції повинні бути ідентичними до учасників конфлікту і за змогою однаковими у всіх конфліктних ситуаціях, щоб не спровокувати нових конфліктів. Застосовуючи санкції варто зважати на персоналії порушників. Якщо жоден з учасників не становить для спільноти особливої цінності, тобто не належить до корисних груп учасників, то до усіх учасників конфлікту застосовуються навіть найжорсткіші санкції – блокування доступу до спільноти. Якщо у конфлікті брав участь корисний учасник спільноти, є сенс зробити учасникам конфлікту попередження і надалі пильно стежити за учасниками, які стали ініціаторами конфлікту (помітити їх

як проблемних), за необхідності застосувати жорсткіші санкції.

10. Приділяти більшу увагу дискусіям, у яких беруть участь особи, між якими вже траплявся конфлікт.

11. Брати активну участь в обговореннях тем та «спрямовувати» дискусію, яка починає перетворюватися на конфлікт, в інше русло, знешкоджуючи у такий спосіб конфлікт ще до його виникнення або на початковій стадії розвитку.

Типові ознаки конфліктних ситуацій, які виявляють за допомогою аналізу дискусій онлайн спільноти такі:

1.У дискусіях одне за одним з'являються повідомлення лише двох учасників спільноти.

2.Час між появою нових повідомлень у дискусії надзвичайно малий (порівняно із середньостатистичним).

3.У повідомленнях у дискусії використовується багато цитувань (визначається за наявністю великої кількості тегів цитування «quote»).

4.У повідомленнях у цій дискусії виявлено багато смайликів.

Типові лінгвістичні ознаки конфліктних ситуацій поділяються на два основні підвиди: лексичні (явні) та смислові(приховані).

Побудова алгоритму застосування санкцій до учасників онлайн спільноти

Алгоритм застосування санкцій до учасників онлайн спільноти призначений для обмеження доступу до онлайн спільноти учасників, які негативно впливають на спільноту, тобто якщо порушення правил, вчинені ними, переважають їхню корисність. Алгоритм є логічним продовженням алгоритму відстеження небажаного інформаційного наповнення. Блок-схема алгоритму наведена на рис. 5.

Адміністратор, виявивши небажане інформаційного, наповнення дані про порушення і порушника, за допомогою автоматизованого алгоритму застосування санкцій визначає міру покарання.

Алгоритм складається з таких кроків:

1. Нарахування штрафних балів учаснику залежно від вчиненого порушення.

2. Визначення корисності учасника для спільноти.

3. Порівняння корисності учасника з кількістю його штрафних балів.

4. Якщо кількість штрафних балів менша за корисність учасника, йому надсилається попередження.

Якщо ж кількість штрафних балів не менша за корисність учасника, алгоритм обмежує учаснику доступ до онлайн спільноти.

5. Алгоритм надсилає адміністратору звіт про виконані дії.

K – коефіцієнт штрафу. Цей коефіцієнт призначений для коректного порівняння штрафних балів і корисності учасника. Визначається окремо для кожної онлайн спільноти (для «Форуму рідного міста» коефіцієнт K рівний 100). Цей алгоритм дає змогу автоматизувати процес адміністрування учасників, оперативно блокувати доступ до онлайн спільноти небажаним учасникам. Це сприятиме зменшенню кількості конфліктів та скороченню часу на модерацію спільноти, і, завдяки цьому – зростанню контрольованості спільноти і підвищенню ефективності онлайн спільноти онлайн спільноти. Крім того, він дає змогу уникнути суб'єктивності адміністратора під час прийняття рішень, оскільки спирається на факти порушень.

Рис. 5. Блок-схема алгоритму застосування санкцій до учасників онлайн спільноти

Висновки

Описані у цій статті методи відстеження появи небезпечного контенту для користувачів онлайн спільнот розроблена для підвищення інформаційно-психологічної безпеки онлайн-користу-вачів та уникнення випадків притягнення до адміністративної та кримінальної відповідальності власників, адміністраторів, модераторів та пересічних користувачів онлайн спільнот.

Щоб відстежити появу небезпечного контенту для користувачів онлайн спільнот розроблено такі методи, як методи недопущення, виявлення та розв'язання конфліктів між учасниками онлайн спільноти, метод відстеження появи небажаного інформаційного наповнення та метод застосування санкцій до учасників онлайн спільноти.

Розроблення запропонованих методів гостро потребують адміністратори та модератори онлайн спільнот з жорсткою структурою, строгою ієрархією та своєрідною спеціалізацією (наприклад, закритих корпоративних онлайн спільнот, онлайн спільнот навчальних курсів, вищих навчальних закладів та шкіл, державних, військових та поліцейських установ та ін.).

Література

[1] J. Libbenga, «German court rules moderators liable for forum comments». [Електронний ресурс]. Режим доступу:www.theregister.co.uk/2006/04/21/moderator_liable_for_comments.

[2] Відповідальність власника сайту. Юридичний довідник онлайн. [Електронний ресурс]. Режим доступу: https://legalsos.com.ua/tsyvilnividn

osyny/vidshkoduvannya-shkody/vidpovidalnist-vlasnykasajtu. html

[3] Л. Черняк, «Порталы и жизненные циклы». [Електронний ресурс]. Режим доступу: http://www.osp.ru/os/2002/02/181136.

[4] A. Peleschyshyn, Yu Syerov, S. Fedushko, «Developing algorithm of registration and validation of personal data on web-community member», *Journal of the Lviv Polytechnic National University: Computer Science and Information Technology*, Lviv, (ed.) NU LP, №686, pp. 238-244, 2010.

[5] А. Пелещишин, Р. Кравець, Ю. Сєров, С. Федушко, «Методи відстеження появи небажаного інформаційного наповнення Веб-форуму», *Вісник Національний університет «Львівська політехніка». Серія Інформаційні системи та мережі*, Львів, № 689, с. 303-312, 2010.

[6] Ю. Сєров, Р. Кравець, А. Пелещишин, «Методи аналізу ефективності Веб-форумів. Інформаційні системи та мережі», *Вісник Національного університету «Львівська політехніка»*, № 653, с.197–206, 2009.

[7] B. Choi., I. Lee, «Trust in open versus closed social media: The relative influence of user- and marketer-generated content in social network services on customer trust», *Telematics and Informatics*. T. 34, № 5, p. 550-559, 2017.

[8] S. Papadopoulos., Y. Kompatsiaris., A. Vakali. P. Spyridonos, «Community detection in social media», *Data Mining and Knowledge Discovery*, № 24, 3 (2012), pp. 515–554, 2012.

[9] L. Yang, C. Tang, H. Wang, H. Tang, «Multi-path Routing Policy for Content Distribution in Content Network», *Ksii Transactions on Internet and Information Systems,* T. 11, № 5, pp. 2379-2397, 2017.

[10] S. Fedushko, Yu. Syerov., R. Korzh, «Validation of the user accounts personal data of online academic community», *Proceedings of the XIIIth International Conference «Modern Problems of Radio Engineering, Telecommunications and Computer Science» (TCSET'2016)*, pp. 863-866, 2016.

[11] S. Fedushko, Yu. Syerov, A. Peleschyshyn, R. Korzh, «Determination of the account personal data adequacy of web-community member», *International Journal of Computer Science and Business Informatics*, Vol. 15, No. 1, p.1-12, January, 2015. [Online]. Available at: http://ijcsbi.org/index.php/ijcsbi/article/view/506/144

[12] С. Федушко, Ю. Сєров, «Програмний комплекс верифікації персональних даних веб-учасника», *Інформація, комунікація, суспільство: матеріали IV Міжнародної наукової конференції ІКС-2015,* Львів, с.58-59, 2015.

[13] P. Paganini, «The business of Censorship. Golden Shield Project, but not only». [Online]. Available at:: http://securityaffairs.co/wordpress/204/cybercrime/business-of-censorship-golden-shield-project-but-not-only.html

[14] S. Fedushko, N. Bardyn, «Algorithm of the cyber criminals identification», *Global Journal of Engineering, Design & Technology (GJEDT)*, Vol. 2, No. 4(2013), p. 56-62, 2013. [Online]. Available at: http://www.gifre.org/admin/papers/gjedt/ALGORITHMVol%202%284%29-gjedt.pdf

[15] S. H. Liu, T. Forss, «Text Classification Models for Web Content Filtering and Online Safety», *IEEE International Conference on Data Mining Workshop*, p. 961-968, 2015.

[16] Y. Liu, S. H. Yang, «Content Filtering Research Based on Web Community Structure», *Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (Mec)*, p. 1773-1777, 2013.

[17] S. Lovaas, «Web Monitoring and Content Filtering, in Computer Security Handbook, Sixth Edition, John Wiley & Sons», Hoboken, USA, 2013.

[18] D. D. Nguyen, M. Erdmann, T. Takeyoshi, G. Hattori, K. Matsumoto, C. Ono, «Training Multiple Support Vector Machines for Personalized Web Content Filters», *IEICE Transactions on Information and Systems*, T. E96D, № 11, pp. 2376-2384, 2013.

[19] S. Sathish., A. Patankar, N. Priyodit, N. Neema, «Enabling Custom Application Content through Semantic Web Filters», *2015 International Conference on Science in Information Technology*, p. 241-246, 2015.

[20] Community Standards. [Online]. Available at:: https://www.facebook.com/communitystandards.

[21] Twitter's Misbegotten Censorship. [Online]. Available at: https://www.theatlantic.com/politics/archive/2016/11/twitter-censorship-will-only-empower-the-altright/507 929/.

[22] Ad Policy: Hate content, sensitive topics, and violence. Twitter Help Center. [Online]. Available at: https:// support.twitter.com/articles/20170425

[23] Facebook and Twitter pledge to remove hate speech within 24 hours. CNN Tech. [Online]. Available at: http://money.cnn.com/2016/05/31/technology/ha tespeech-facebook-twitter-eu/index.html.

[24] R.Suvorov, I. Sochenkov, I. Tikhomirov, «Training Datasets Collection and Evaluation of Feature Selection Methods for Web Content Filtering», *Artificial Intelligence: Methodology, Systems, and Applications*, Cham, pp. 129-138, 2014.

[25] Web Forum «File Sharing Talk». When the **** did we get a wordfilter? [Online]. Available at: https://filesharingtalk.com/threads/88125-When-the-****-did-we-get-a-wordfilter.

[26] Wordfilter – TheInfoList. [Online]. Available at: http://www.theinfolist.com/php/SummaryGet.php?FindGo=Wordfilter

[27] R. Howard, «HOW TO: Manage a Sustainable Online Community». [Online]. Available at: http://mashable.com/2010/07/30/sustainable-online-community.

[28] Ю.О. Сєров, А.М. Пелещишин, К.О. Слобода, «Аналіз комунікативних процесів у Веб-спільнотах середовища Веб 2.0.», *Східно-Європейський журнал передових технологій,* Х., № 1/2 (37), с. 38-41, 2009.

[29] Yu. Syerov, R. Kravets, «Software for determination of behavior patterns of web-forum members», *Міжнародний науково-технічний журнал «Комп'ютинг»*, Т. 8, № 2, с.119-125, 2009.

[30] С.С. Федушко, «Аналіз архітектури та сучасних тенденцій розвитку віртуальних спіль-

нот», *Вісник НУ ЛП: Інформаційні системи та мережі*, № 699, Львів, с. 362-375, 2011.

[31] А.М. Пелещишин., О.Р. Трач, «Визначення етапів життєвого циклу віртуальної спільноти. Управління розвитком складних систем», *Київський нац. університет будівництва і архітектури*, К., с. 133-137, 2014.

[32] Пелещишин А., Сєров Ю. , Федушко С., «Розроблення алгоритму реєстрації та валідації персональних даних учасників Веб-спільноти», *Вісник Національного університету «Львівська політехніка»: Комп'ютерні науки та інформаційні технології*, Львів, №686, с. 238-244, 2010.

УДК 004.773.2 (045)

Серов Ю. О. Отслеживание появления опасного контента онлайн сообществ как ключевой аспект информационно-психологической безопасности онлайн-пользователей

Аннотация. *В статье предложено решение актуальной задачи разработки методов отслеживания появления опасного контента для пользователей онлайн сообществ с целью повышения их информационно-психологической безопасности. Внедрение полученных результатов в работу онлайн сообщества позволяют избежать случаев привлечения к административной и уголовной ответственности владельцев, администраторов, модераторов и рядовых пользователей онлайн сообществ. Методы недопущения, выявления и разрешения конфликтов между участниками онлайн сообщества, метод отслеживания появления нежелательного информационного наполнения и метод применения санкций к участникам онлайн сообщества разработано с целью отслеживания появления опасного контента для онлайн сообществ. Предложенные методы существенно повысят информационно-психологическую безопасность пользователям, администраторам и модераторами онлайн сообществ с жесткой структурой, строгой иерархией и своеобразной специализацией, а именно закрытых корпоративных онлайн сообществ, онлайн сообществ учебных курсов, высших учебных заведений и школ, государственных, военных и полицейских учреждений.*

Ключевые слова: *онлайн сообщество, опасный контент, пользователь, отслеживание угроз, информационное наполнение, фильтрация запрещенного контента.*

Syerov Yu. Tracking of dangerous content of online communities as a key aspect of information-psychological security of online users

Abstract. *The article proposed solution actual problem of developing methods for tracking of dangerous content for users of online communities to improve their information-psychological security. Implementation of the received results in online communities work allow owners, administrators, moderators and users of online communities to avoid cases of instituting administrative and criminal liability. Methods of prevention, detection and resolution of conflicts between online community members, method of tracking unwanted content and method of applying sanctions to member's online community are developed to tracking dangerous content to online communities' users. The proposed methods will significantly increase the information and psychological safety of users, administrators and moderators of online communities with a inflexible structure, strict hierarchy and peculiar specialization, as namely, closed corporate online communities, online communities of training courses, universities and schools, government, military and police agencies.*

Key words: *online community, dangerous content, user tracking threats, filtering of prohibited content.*

Отримано *6 червня 2017 року*, затверджено редколегією *4 липня 2017 року*