УДК 81'11+32+33: 004.9

С.С. Федушко, аспірант, асистент,
Національний університет «Львівська політехніка», м. Львів, Україна
felomia@gmail.com

Застосування апарату нечітких множин для класифікації учасників віртуальної спільноти

В статті розглянуто методи визначення рівня освіченості учасників віртуальних спільнот на основі теорії нечітких множин. Створено модель класифікації учасників віртуальних спільнот за рівнем їх освіченості. На основі алгоритму визначення рівня освіченості учасника веб-спільноти розроблено програмний засіб з метою автоматизованої перевірки грамотності та класифікації учасників веб-спільнот за цим критерієм.

Ключові слова: алгоритм, віртуальна спільнота, рівень освіченості, лінгвістична змінна.

Вступ

Адміністратори багатьох спільнот вимагають від учасників віртуальних спільнот дотримуватись певних конвенцій веб-спілкування. Конвенції веб-спілкування залежать від поставленої мети завдань та прогнозованих власниками сценаріїв розвитку віртуальної спільноти (ВС). Важливим фактором у можливості участі у віртуальній спільноті є високий рівень грамотності веб-учасника, рівень вищої освіти та навиків.

Метод відсіювання недостатньо грамотних та з низьким рівнем освіченості учасників віртуальної спільноти значно зменшить затрати часу, фінансів та зусиль адміністраторів та модераторів на модерування віртуальною спільнотою. Для відсіювання неграмотних учасників віртуальної спільноти потрібно всіх учасників класифікувати за рівнем грамотності, що дасть зможу визначити ймовірний рівень освіти кожного учасника віртуальної спільноти.

Аналіз основних досліджень і публікацій

В дослідженнях віртуальних спільнот у WWW науковці виокремлюють багато напрямів і у зв'язку з появою нових тенденцій розвитку віртуальних спільнот сфера досліджень збільшується буквально кожного дня.

Аналіз сучасних праць науковців виявив, що в останні роки з'явилося багато публікацій, які присвячено вивченню особливостей функціонування віртуальних спільнот, способів формування мережної ідентичності веб-особистості.

На сьогоднішньому етапі розвитку віртуальних товариств науковці [1, 2, 4, 6] розглядають соціолого-психологічні аспекти комунікації людей (K. Popper, A. Podgorecki, J. Karpinski); моніторинг дій користувачів (B. Mobasher, S. Chakrabarti); формування спільнот та управління ними; організація опрацювання інформаційного наповнення (W. Cunningham, T. O'Reilly); позиціонування інформаційного наповнення (И. Ашманов); підвищення якості інформаційного наповнення (T.Berners-Lee); визначення критеріїв якості інформаційного наповнення (M. Parker).

Цей огляд, звичайно, не охоплюють всі дослідження, які відбулись в останні кілька років, однак, він дає уявлення з-поміж великої кількості про найважливіші роботи, які формують основу для області, і дослідження, які розширили нові горизонти в області комп'ютерно-опосередкованої комунікації.

В українському сегменті Інтернету, Укрнеті, ці напрямки досліджень, попри важливість, почали науковці розглядати та аналізувати недавно, незважаючи на його доволі швидкий розвиток. Втім, проаналізувавши існуючі дослідження та публікації, варто зауважити, що кількість фундаментальних досліджень у цій області невелика, проте науковий інтерес зростає і є вже доволі вагомі результати досліджень.

Методи побудови інформаційного суспільства, соціальні мережі в WWW та інформаційні технології соціальних комунікацій досліджує у своїх роботах Пелещишин А.М. [8-10].

Серед відомих типів віртуальних спільнот, веб-форуми є одним із найпотужніших та найпопулярніших сервісів WWW, які призначені для організації спілкування користувачів мережі. Веб-форуми є унікальним джерелом інформації, місцем накопичення великих обсягів важливої, цінної інформації та знань, рушієм різноманітних комерційних та суспільних проектів.

Веб-форуми є ефективним інструментом комунікації викладачів і учнів у системах дистанційного навчання та об'єктом досліджень конфліктів між учасниками віртуальних спільнот.

Активність учасника веб-форуму досліджує Сєров Ю.О. [8, 9, 11, 12] і визначав активність користувача веб-форуму, як характеристику учасника веб-спільноти, яка визначає його здатність створювати інформаційне наповнення і визначається кількістю повідомлень учасника та нових дискусій, які він пропонує розпочати.

Всі ці напрямки досліджень, попри важливість, розглянуті та проаналізовані доволі поверхнево. Все вище сказане обумовлює актуальність досліджень віртуальних спільнот.

Отже, постає доволі актуальні проблеми управління віртуальними спільнотами і її вирішення полягає у створенні нових підходів та методів, які представлені у цьому дослідженні.

Мета та основні завдання статті

Мета дослідження є розроблення методу класифікації учасників веб-спільноти за рівнем освіченості (грамотності створення ними інформаційного наповнення спільноти). А саме, алгоритму визначення рівня освіченості учасника віртуальної спільноти.

Для досягнення мети були поставлені такі завдання:

- аналіз існуючих досліджень віртуальних спільнот;
- створити модель класифікації учасників веб-спільнот за рівнем освіченості на основі теорії нечітких множин;
- розробити алгоритм визначення рівня освіченості учасника віртуальної спільноти.

Модель класифікації учасників веб-спільнот за рівнем освіченості

У цьому дослідженні задачі автоматизації виявлення помилок у тексті розглядаються тільки для випадків символьних помилок. Проведений науковцями, аналіз помилок та консолідація цих даних дозволяє запропонувати типологію помилок. Отож, наведемо типові помилки у інтернет-комунікації:

1. Заміна літери на подібну по звучанню:*"білий грип – білий гриб";*
2. Пропуск літер (найчастіше голосних) : *"цього – цьго", "зосереджуються –зосереджються";*
3. Подвоєння літер (найчастіше приголосних): *"частотність – чаастотність", "головне – гооловне";*
4. Перестановка літер: найчастіше сусідніх клавіш).(*"правильно – рпавильно";*
5. Випадкове вставлення зайвої літери: *"політика – поліътика";*
6. Вставка зайвих пробілів перед чи після слів: *"ідентичне_ _ питання;*
7. Відсутність пробілу чи дефісу: *"Будь який", "недостатня увага –недостатняувага";*
8. Схожість написання цифр і букв (ч-4, про-0, з-3):*"4естно – чесно";*
9. Ідентичне написання букв в різних розкладках. (*"XP0M0C0MA – хромосома");*
10. Букви і символи в різних розкладках. (*"<лізнец – близнюк");*
11. Переплутування і зміщення рук при сліпому наборі літер: *"інвнжае – телефон";*
12. Переклад транслітерації на українське написання: *"Kartinki – картинки";*
13. Виправлення неправильної розкладки клавіатури як для цілого, так і для частини слова: *«jlyjrkfccybrb – однокласником».*

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

Для зручності аналізу множина лінгвістичних ознак учасника віртуальної спільноти, які вказують на рівень освіти, описуємо наступним чином:

$$Edu(U_i) = \left(Edu(U_i)_j\right)_{j=1}^{N^{Edu}} \quad (1)$$

де $\left(Edu(U_i)_j\right)_{j=1}^{N^{Edu}}$ - множина лінгвістичних ознак, які вказують на рівень освіти, учасника віртуальної спільноти;

N^{Edu} - кількість ознак, які вказують на рівень освіти, конкретного учасника віртуальної спільноти.

Для пошуку та корекції цієї множини помилок розроблено науковцями багато алгоритмів як для англійської, так і для української мови, хоч і не настільки ґрунтовно. Зважаючи на вище сказане, розроблення нового автоматизованого засобу для пошуку і кореляції помилок у веб-контенті не є вирішальним.

Оскільки, оптимальним вирішенням цієї задачі є існуючий ефективно функціонуючий автоматизований засіб [3, 5], який полягає в аналізі тексту, фільтруванні слів, відборі слів з помилками та їх корекції.

Визначення рівня освіченості учасників віртуальної спільноти здійснюється методом обчислення значень лінгвістичної змінної та міри належності.

При розробленні методу визначення рівня освіченості учасника віртуальної спільноти введемо лінгвістичну змінну: "Рівень освіченості".

Використання лінгвістичної змінної обумовлене тим, що: цей метод є концептуально простіший для розуміння; моделює нелінійні функції довільної складності; враховується досвід фахівців-експертів; нечітка логіка заснована на природній мові людського спілкування.

Формально лінгвістичну змінну опишемо, як кортеж наступних елементів $\langle \beta, T, X, M \rangle$, де:

- назва лінгвістичної змінної (β) – "Рівень освіченості";
- базову терм-множину лінгвістичної змінної визначемо так:

T={"високий", "середній", "низький", "дуже низький"} $= \{G_1, G_2, G_3, G_4\}$;

- $X = [0; Q(Mistake)]$, де $Q(Mistake)$ – рівень грамотності;
- M – сукупність мір належності нечітких змінних, які є значеннями лінгвістичної змінної "Рівень освіченості" .

Запишемо цю терм-множину в символічній формі $T = \{HL, AL, LL, ML\}$.

Функція належності термів "високий", "середній", "низький" та "дуже низький" лінгвістичною змінною "Рівень освіченості" показані на Рисунку 1.

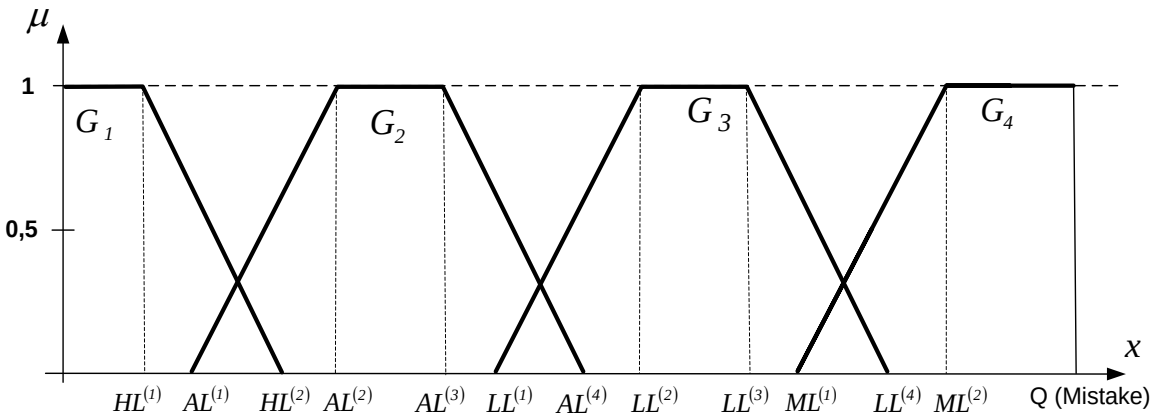

Рисунок 1 – Функція належності нечітких множин:
G1– «високий», G2 – «середній», G3 – «низький», G4 – «дуже низький».

де $HL^{(1)}, HL^{(2)}, AL^{(1)}, AL^{(2)}, AL^{(3)}, AL^{(4)}, LL^{(1)}, LL^{(2)}, LL^{(3)}, LL^{(4)}, ML^{(1)}, ML^{(2)}$–параметри, що пропорційні до $Q(Mistake)$. Ці параметри задає експерт, $HL^{(1)} \le HL^{(2)} \le AL^{(1)} \le AL^{(2)} \le AL^{(3)} \le AL^{(4)} \le LL^{(1)} \le LL^{(2)} \le LL^{(3)} \le LL^{(4)} \le ML^{(1)} \le ML^{(2)} < 1$.

Належності нечітких множин, які використовуються у функції (4 - 7), експерти визначають для кожного учасника віртуальної спільноти за формулою (2). Формулу (2) виведено на основі методів оцінювання рівня знань учнів та студентів у навчальному закладі [7].

$$Q(Mistake) = \begin{cases} HL, k \le c_1 \\ AL, c_1 < k \le c_2 \\ LL, c_2 < k \le c_3 \\ ML, c_3 < k \le c_4 \end{cases} \qquad (2)$$

де k – коефіцієнт грамотності.

Коефіцієнт грамотності обчислюється згідно з формулою (3):

$$k = \frac{n_{mistakes}}{N_{words}}, \qquad (3)$$

де $n_{mistakes}$ – кількість помилок у дописі учасника веб-спільноти; N_{words} – загальна кількість слів у дописі учасника веб-спільноти.

Змінні $n_{mistakes}$ та N_{words} виміряються в інтервалі дійсних чисел.

Відповідно до сучасних методі оцінювання рівня знань змінна с приймає такі значення: $c_1 = 0{,}02;\ c_2 = 0{,}05; c_3 = 0{,}1; c_4 = 1$.

Як правило, великим є значення коефіцієнту грамотності при великій кількості помилок у інформаційному сліді учасника ВС.

Запишемо функції належності для соціально-демографічної характеристики освіченості (4)–(7) учасників віртуальної спільноти.

$$\mu_{HL}(Edu) = \begin{cases} 1, 0 \le Education_{Level}(User_i) < HL^{(1)} \\ \dfrac{HL^{(2)} - Education_{Level}(User_i)}{HL^{(2)} - HL^{(1)}}, HL^{(1)} \le Education_{Level}(User_i) \le HL^{(2)} \end{cases} \qquad (4)$$

$$\mu_{AL}(Edu)=\begin{cases}\dfrac{Education_{Level}(User_i)-AL^{(1)}}{AL^{(2)}-AL^{(1)}}, AL^{(1)}\le Education_{Level}(User_i)<AL^{(2)}\\ 1, AL^{(2)}\le Education_{Level}(User_i)\le AL^{(3)}\\ \dfrac{AL^{(4)}-Education_{Level}(User_i)}{AL^{(4)}-AL^{(3)}}, AL^{(3)}<Education_{Level}(User_i)<AL^{(4)}\end{cases} \quad (5)$$

$$\mu_{LL}(Edu)=\begin{cases}\dfrac{Education_{Level}(User_i)-LL^{(1)}}{LL^{(2)}-LL^{(1)}}, LL^{(1)}\le Education_{Level}(User_i)<LL^{(2)}\\ 1, LL^{(2)}\le Education_{Level}(User_i)\le LL^{(3)}\\ \dfrac{LL^{(4)}-Education_{Level}(User_i)}{LL^{(4)}-LL^{(3)}}, LL^{(3)}<Education_{Level}(User_i)<LL^{(4)}\end{cases} \quad (6)$$

$$\mu_{ML}(Edu)=\begin{cases}\dfrac{Education_{Level}(User_i)-ML^{(1)}}{ML^{(2)}-ML^{(1)}}, ML^{(1)}\le Education_{Level}(User_i)<ML^{(2)}\\ 1, ML^{(2)}\le Education_{Level}(User_i)\le 1\end{cases} \quad (7)$$

де $Education_{Level}(User_i)$ – рівень освіченості i-го учасника віртуальної спільноти.

Таким чином, у цій частині роботи реалізовано поняття лінгвістичної змінної, що дало змогу представити рівень освіченості користувачів веб-спільноти за допомогою нечітких значень відповідно до терм-множини та описано метод визначення міри належності учасників віртуальних спільнот до нечітких множин.

Алгоритм визначення рівня освіченості учасника віртуальної спільноти

Головна мета алгоритму – визначення рівня грамотності учасника веб-форумів методом розпізнання слів написаних з помилками.

Алгоритм полягає у 16 етапах.

Основними кроками алгоритму є:

2. Формування та аналіз інформаційного сліду учасника віртуальної спільноти.

Інформаційний слід – множина всіх даних учасника віртуальної спільноти та результати його комунікативної діяльності – створене ним інформаційне наповнення.

Поняття "інформаційний слід", описано за допомогою формул (8 та 9):

$$InfTrack(U_i)=\left\langle\begin{matrix}Content(U_i),\\ PersonalData(U_i)\end{matrix}\right\rangle, \quad (8)$$

Складовими інформаційного сліду є: інформаційне наповнення – *Content(U_i)*, яке створене учасником віртуальної спільноти, та персональні дані – *PersonalData(U_i)*.

Інформаційне наповнення визначається кортежем, а сааме, такими підмножинами, як дискусії, опитування та дописи:

$$Content(U_i)=\left\langle\begin{matrix}Poll(U_i), Post(U_i),\\ Thread(U_i),\end{matrix}\right\rangle, \quad (9)$$

де $Thread(U_i)=\{Thread_j(U_i)\}_{j=1}^{N_i^{(UThead)}}$ – множина дискусій, створених учасником віртуальної спільноти U_i; $N_i^{(UThead)}$ – кількість таких дискусій.

$Poll(U_i)=\{Poll_j(U_i)\}_{j=1}^{N_i^{(UPoll)}}$ – множина опитувань, створених учасником віртуальної спільноти U_i; $N_i^{(UPoll)}$ – кількість таких опитувань.

$Post(U_i)=\{Post_j(U_i)\}_{j=1}^{N_i^{(UPos)}}$ – множина дописів учасника ВС U_i; $N_i^{(UPos)}$ –кількість дописів учасника веб-спільноти U_i.

3. Знайдення та виправлення помилок. Аналіз відбувається на основі існуючого алгоритму пошуку та кореляції помилок.

4. Класифікація учасників віртуальної спільнот відповідно до їхнього рівня грамотності та запис їх у відповідні бази даних.

5. Надсилання результатів на електронну скриньку модераторів та адміністраторів спільноти.

Отже, з метою спрощення управління та підвищення ефективності діяльності віртуальних спільнот розроблено основу для комп'ютерно-лінгвістичного методу перевірки інформаційного сліду учасника віртуальної спільноти, який базується на лінгвістичному аналізі інформаційного наповнення веб-спільноти.

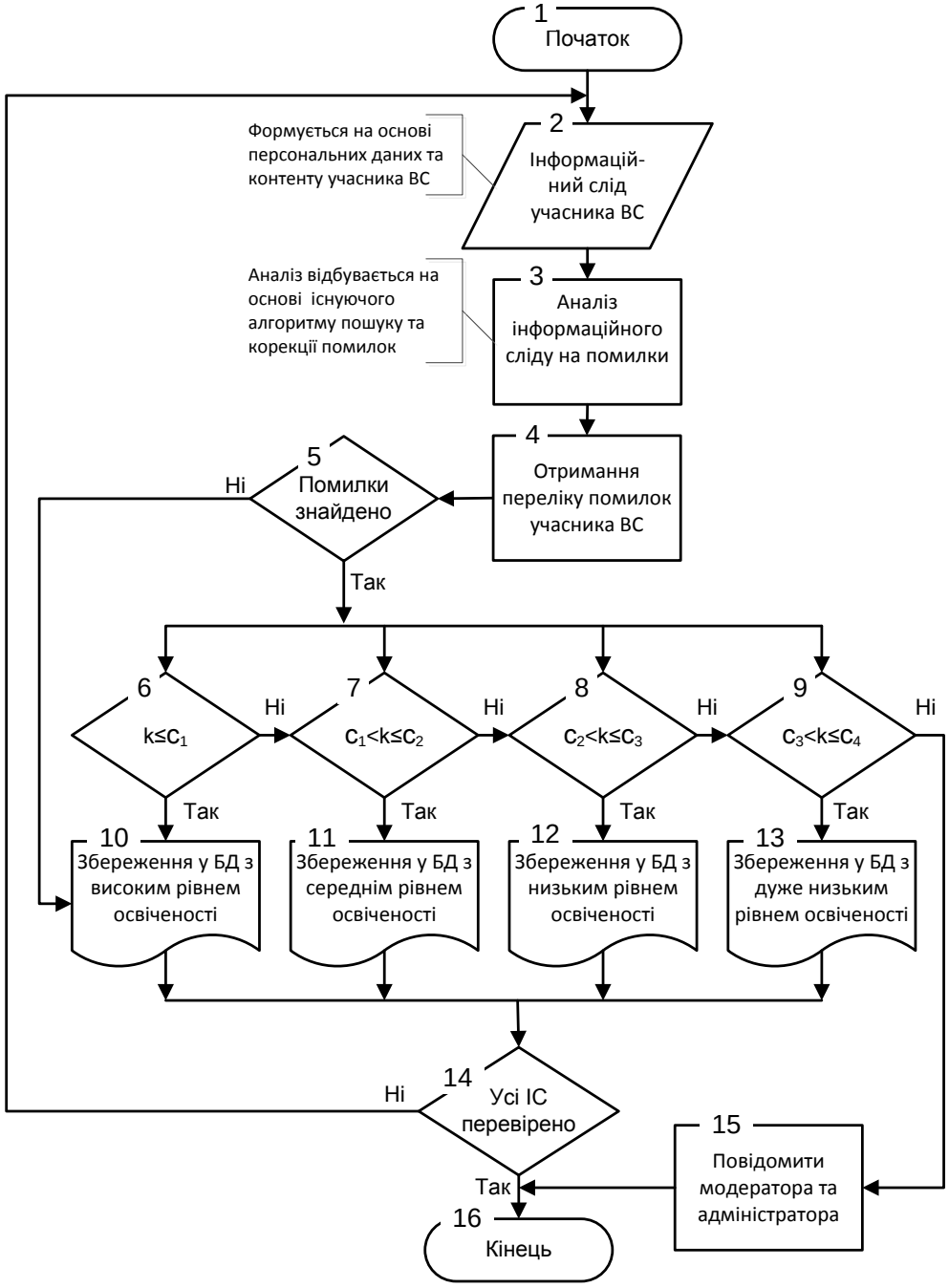

Рисунок 2 - Блок-схема алгоритму визначення рівня освіченості учасника веб-спільноти

Висновок

Розроблення методу підвищення ефективності функціонування та модерування віртуальними спільнотами шляхом класифікації учасників цих спільнот за рівнем грамотності представлено у цьому дослідженні. Ця класифікація показана у алгоритмі визначення рівня освіченості (високий, середній, низький та дуже низький рівень грамотності) користувача веб-спільноти, що є ефективним засобом для більш структурного та поступового моніторення комунікативної поведінки кожного учасника спільноти та відсіювання небажаних учасників.

Програмний засіб створений на основі цього алгоритму є ефективним вирішенням актуальних задач у управлінні веб-спільнотами, системами електронного урядування, дистанційного навчання та інтернет-маркетинговій сфері.

Список літератури

1. Carr N. The amorality of Web 2.0. [Электронный ресурс] / N. Carr // Rough type. - 2005. – Режим доступа: http://www.roughtype.com/archives/2005/10 /the_amorality_o.php.
2. Croll A. Complete Web Monitoring: Watching Your Visitors, Performance, Communities, and Competitors / A. Croll, S. Power. – O'Reilly Media, 2009. – 672 p.
3. Inkpen D. Unsupervised approaches to text correction for English and Romanian / D. Inkpen, A. Islam ; School of Information Technology and Engineering. - Ottawa. –PP. 272-287.
4. Kravets R. Web forum member behaviour modeling and classifying based on fuzzy sets / R. Kravets, A. Peleschyshyn, Yu. Syerov // Proceedings of the International Conference of Computer Science and Information Technologies «CSIT'2007». – Lviv, 2007. – PP. 279–280.
5. Kukich K. Techniques for automatically correcting words in text / K. Kukich // ACM Computing Surveys (CSUR). - 1992. - Volume 24, issue 4. - PP. 377-439.
6. Zhang Y. Web Communities: Analysis and Construction / Y. Zhang, J.Xu Yu, J.Hou, 2005. – 187 p.
7. Наказ Міністерства освіти і науки України "Про затвердження критеріїв оцінювання навчальних досягнень учнів у системі загальної середньої освіти" від 05.05.2008 N 371.
8. Процеси управління інтерактивними соціальними комунікаціями в умовах розвитку інформаційного суспільства: монографія / [А. М. Пелещишин, Ю. О. Сєров, О. Л. Березко та ін.]; за заг. ред. А. М. Пелещишина. – Львів: Видавництво Львівської політехніки, 2012. – 368 с.
9. Пелещишин А. Методи відстеження появи небажаного інформаційного наповнення Веб-форуму / А. Пелещишин, Ю. Сєров, С. Федушко // Вісник НУ "Львівська політехніка": Інформаційні системи та мережі. – 2010. – №689. – С. 303-312.
10. Пелещишин А. М. Оптимізація форумів та інших форм спільнот користувачів WWW / А. М. Пелещишин // Інформаційні системи та мережі: Вісник НУ "ЛП". – 2004. – № 519. – С. 275-284.
11. Сєров Ю. Методи аналізу функціонування Веб-форумів / Ю. Сєров, Р. Кравець, В. Сівокозов // Proceedings of the Third International Conference of Young Scientists on Computer Science and Engineering (CSE'2009). – Lviv, 2009. – С. 84–86.
12. Сєров Ю. О. Методи аналізу ефективності веб-форумів / Ю. О. Сєров, Р. Б. Кравець, А. М. Пелещишин // Інформаційні системи та мережі: Вісник НУ «ЛП». – 2009. – № 653. – С. 197-206.

Надійшла до редакції 20.05.2013

С.С. ФЕДУШКО
Национальный университет «Львовская политехника»

ПРИМЕНЕНИЕ АППАРАТА НЕЧЕТКИХ МНОЖЕСТВ ДЛЯ КЛАССИФИКАЦИИ УЧАСТНИКОВ ВИРТУАЛЬНОГО СООБЩЕСТВА

В статье рассмотрены методы определения уровня образованности участников виртуальных сообществ на основе теории нечетких множеств. Создана модель классификации участников виртуальных сообществ по уровню их образованности. На основе алгоритма определения уровня образованности участника веб-сообщества разработано программное средство с целью автоматизированной проверки грамотности и классификации участников веб-сообществ по этому критерию.

Ключевые слова: алгоритм, виртуальное сообщество, уровень грамотности, нечеткие множества.

S. S. FEDUSHKO
Lviv Polytechnic National University

CLASSIFICATION OF VIRTUAL COMMUNITIES MEMBERS BASED ON FUZZY SETS

This article considers the current problem of investigation and development of the method of web-members' classification. The aim of this method is to check the correctness of the web-community members' content formation that substantially effects the efficiency of virtual communities functioning and to improve the registration and moderating processes in the online community. The set of frequently used mistakes in web-community members' posts is formed to check the grammatical correctness of web community members, which form based on analysis of web-communities members' information tracks. The algorithm for determining the educational level of web-community member helps to classify web-community members by level of education of the web-members. The result of this algorithm is the classification of web-forum members according to educational level (high, medium, low and very low). Thus, this method can be of essential help to administrators and moderators to monitor users of the web-community. The paper presents a new approach to developing web-member classification method based on fuzzy sets. These issues have the greatest influence on efficiency rise of virtual communities functioning and the level of web-members education. The solution of these problems is possible by using computer-linguistic analysis of web-members' posts.

Keywords: algorithm, virtual community, educational level, fuzzy sets.