УДК 004.773.2

Р.Б. Кравець, Ю.О. Сєров
Національний університет “Львівська політехніка”,
кафедра інформаційних систем і мереж

ПОБУДОВА ПРОГРАМНОЇ СИСТЕМИ АНАЛІЗУ ПОВЕДІНКИ УЧАСНИКІВ ВЕБ-ФОРУМУ

Розглянуто актуальну проблему проектування та побудови інформаційної системи аналізу поведінки та класифікації учасників Веб-спільноти.

Article consideres actual problem of Web-community members behaviour and classifying information system development.

Постановка проблеми

Учасниками Веб-спільнот (форумів) є люди, які природно володіють певним стилем поведінки, мають притаманні їм риси характеру. Життя Веб-спільнот визначають їх учасники, тому очевидною є необхідність постійного моніторингу Веб-спільноти шляхом дослідження та аналізу поведінки учасників Веб-спільноти [1, 2].

Дослідження та аналіз поведінки учасників Веб-спільноти є складним обчислювальним процесом, тому створення інформаційної системи аналізу поведінки учасників Веб-спільноти, яка б органічно інтегрувалася в систему управління інформаційним наповненням (СУІН) форуму, є важливою та актуальною задачею.

Учасники визначають спосіб життя Веб-спільноти, тому власникам Веб-спільнот необхідно володіти програмними засобами аналізу поведінки учасників Веб-спільноти, які допоможуть ефективніше здійснювати модерування та адміністрування.

Аналіз останніх досліджень

Функціонування сайту Веб-спільноти неможливе без програмного комплексу, який би забезпечував виконання базового набору функцій, необхідних для реєстрації учасників, опублікування контенту, управління контентом і учасниками. Сьогодні існують програмні комплекси — системи управління інформаційним наповненням (СУІН). Такі системи мають типову структуру схеми даних і володіють усіма необхідними функціями, які забезпечують функціонування Інтернет-форуму. Тому більшість Веб-форумів функціонують на базі масових популярних СУІН. Такий стан речей, а саме існування типових інформаційних систем Веб-форумів, дає змогу Інтернет-користувачам та учасникам Веб-спільнот швидко звикнути до інтерфейсів та опанувати функціональні можливості цих інформаційних систем. Також це дає змогу розробити для власників та адміністраторів інформаційну підсистему аналізу функціонування та поведінки учасників Веб-спільноти.

Серед масових популярних СУІН є як комерційні продукти, так і такі, що поширюються безкоштовно.

До найпопулярніших на сьогодні СУІН належать vBulletin, Invision, phpBB, XMB.

Однак попри добру функціональність усі сучасні СУІН Веб-спільнот не забезпечують власникам та адміністраторам можливостей аналізувати життя спільноти та поведінку її учасників. СУІН надають лише найпростішу інформацію: про кількість повідомлень, дискусій, нових

учасників, які з'явилися за останню добу, кількість учасників, які відвідали форум тощо. Система не надає детальнішу та складнішу інформацію про динаміку накопичення інформаційного наповнення, про динаміку приросту учасників Веб-спільноти, про корисність учасників. Тому створення інформаційної системи аналізу поведінки учасників, яка може аналізувати діяльність учасника спільноти за весь чи за певний період його участи у Веб-спільноті, класифікувати учасників, давати інтелектуальні поради щодо адміністрування Веб-спільноти, є важливою та актуальною задачею.

Цілі статті

Завданням цієї статті є опис етапів побудови інформаційної системи аналізу поведінки учасників Веб-спільнот.

Перед тим, як почати етап проектування системи, потрібно проаналізувати процес аналізу діяльності учасників Веб-форумів, а також структуру існуючих СУІН Веб-спільнот для того, щоб побудувати інформаційну систему, котра б роз'язувала усі задачі аналізу поведінки учасників Веб-спільнот, а також органічно вписувалася в існуючі системи управління Веб-форумів. На основі проведеного аналізу потрібно спроектувати систему та розробити її прототип.

Отже, цілями цієї статті є:

- аналіз структур існуючих СУІН Веб-форумів;
- дослідження процесу аналізу поведінки учасників Веб-спільнот;
- побудова схеми даних інформаційної системи аналізу поведінки учасників Веб-спільноти;
- програмна реалізація алгоритмів аналізу поведінки учасників Веб-форумів;
- апробація функціонування інформаційної системи на основі даних Веб-спільноти Форум Рідного Міста.

Основний матеріал

Структура та інформаційна модель Веб-форуму

Основні об'єкти Веб-форуму

Основними об'єктами Веб-спільноти є учасники та інформаційне наповнення, яке вони створюють: дискусії та повідомлення.

Учасник Веб-спільноти — людина або інтелектуальний агент (бот), який відвідує сайт Веб-спільноти, читає (зчитує) чи публікує інформацію у вигляді дискусій та повідомлень.

З погляду прав та повноважень, учасники Веб-спільноти належать до одного з класів: незареєстровані учасники (гості), зареєстровані учасники, модератори, адміністратори.

Інформаційна структура Веб-форуму — це ієрархія підфорумів та повідомлень, це дерево з вузлами двох типів: підфоруми і дискусії (рис. 1).

Підфорум – це множина підфорумів нижчого рівня та дискусій.

Дискусія – це множина повідомлень, створених учасниками форуму. Дискусії можуть розміщуватися на будь-якому рівні структури, однак найчастіше – лише на найнижчому.

Кожне повідомлення на форумі може належати до одного з таких типів:

Початок опитування – формулювання опитування із зазначенням теми опитування та варіантами відповідей. Тут важливим є таке формування множини варіантів відповідей, щоб вона була повною (не існувало інших варіантів відповідей).

Початок дискусії – перше повідомлення дискусії, яке ініціює обговорення.

Відповідь — кожне повідомлення в дискусії, яке не є початком дискусії, є відповіддю. Відповідь може містити цитування попередніх повідомлень.

Усе інформаційне наповнення створюється учасниками Веб-форуму. Створене інформаційне наповнення індексується пошуковими машинами, і тим самим стає доступним для користувачів Інтернету, які потім відвідують сторінки Веб-форуму з шуканою інформацією. Таким чином деякі з них стають новими учасниками Веб-спільноти.

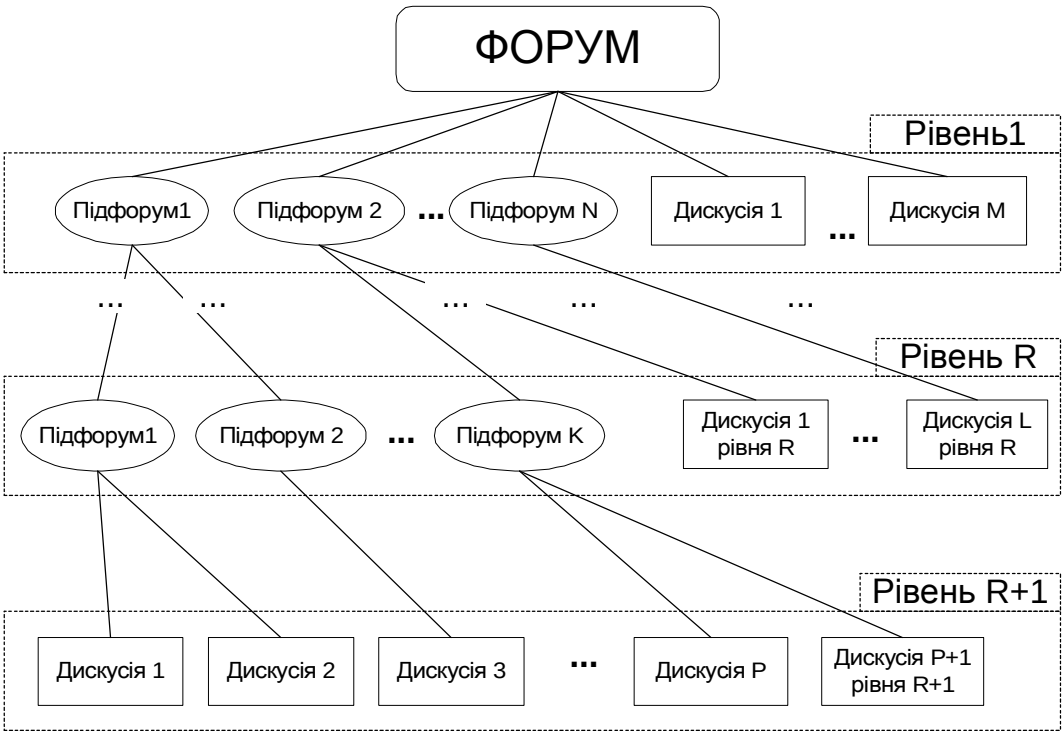

Рис. 1. Деревоподібна структура форуму

Узагальнена інформаційна модель Веб-спільноти

Загалом усі СУІН Веб-форумів мають складну схему даних, яка складається з великої кількості таблиць. Однак значна частина цих таблиць є довідниками, які виконують допоміжну функцію.

Розглянувши та проаналізувавши існуючі СУІН Веб-форумів, відкинувши таблиці-довідники та інші другорядні сутності, виділимо основні об'єкти Веб-спільноти як інформаційної системи. Узагальнена схема даних, яка забезпечує необхідний набір сутностей для аналізу поведінки учасників Веб-спільнот, має такий вигляд (рис.2).

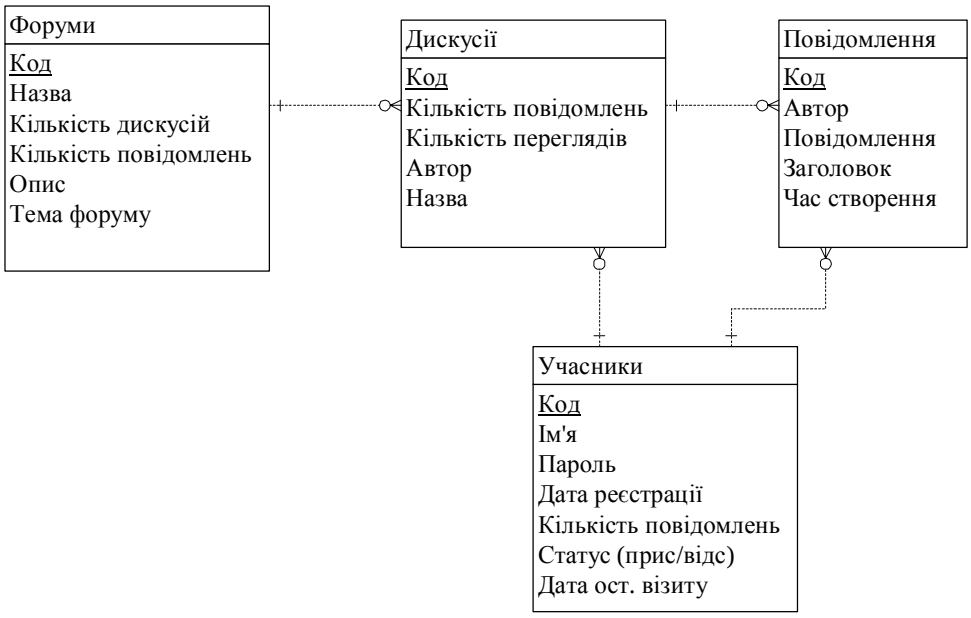

Рис. 2. Узагальнена структура даних "ядра" СУІН Веб-спільноти

Основними складовими такої інформаційної системи, як Веб-спільнота, є **Учасники (Members)** та **Інформаційне наповнення.**

Інформаційне наповнення структурно складається з Форумів (Forums), Дискусій (Threads) та Повідомлень (Posts).

Учасник Веб-спільноти характеризується такими головними параметрами:

Ім'я, Пароль, Персональні дані, Множина повідомлень, Множина розпочатих дискусій, Множина отриманих відгуків, Множина висловлених відгуків.

Інша частина ядра – інформаційне наповнення – представлене трьома сутностями:

Форум характеризується такими параметрами:

Назва, Кількість дискусій (які належать до цього форуму), Кількість повідомлень (які належать до цього форуму), Опис.

Дискусія характеризується такими параметрами:

Назва, Автор, Форум (до якого належить), Кількість повідомлень, Кількість переглядів.

Повідомлення характеризується такими параметрами:

Заголовок, Тіло повідомлення, Автор, Дата, Дискусія (до якого належить), Форум (до якого належить).

Перед тим як безпосередньо перейти до опису інформаційної моделі системи, коротко опишемо поведінку учасників спільноти, їх риси та способи їх класифікації.

Учасникам Веб-спільнот притаманні такі риси поведінки: Активність, Креативність, Атрактивність, Реактивність, Лояльність.

Для моделювання рис учасників спільноти застосовується апарат нечітких множин [2]. Усі перелічені риси учасників спільноти, відносно усього форуму можна визначити, аналізуючи їх поведінку в межах спільноти: активність визначається кількістю контенту, які вони створюють; креативність – якістю контенту і тим, як реагують на нього інші учасники спільноти; атрактивність – кількістю учасників, які реагують на створений контент; реактивність – способом участи в дискусіях, лояльність – реакцією на контент інших учасників.

Програмна реалізація

Узагальнена логічна інформаційна модель Веб-форуму

Наведемо логічну інформаційну модель бази даних інформаційної системи, що містить інформацію, яка аналізується.

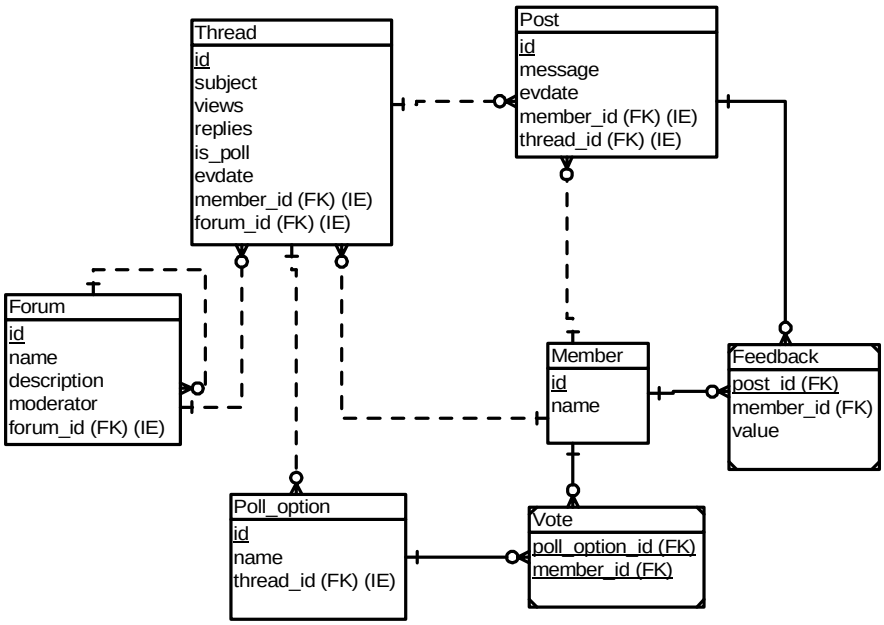

Рис. 3. Інформаційна модель Веб-форуму

Узагальнена інформаційна модель Веб-форуму, зображена на рис. 3, складається з чотирьох сутностей: Форуми, Дискусії, Повідомлення, Учасники. Учасників класифікуємо на основі інформації, що міститься у цій базі даних.

Інформаційна модель даних для класифікації учасників

Важливим етапом створення інформаційної системи аналізу поведінки учасників Веб-спільноти на основі форуму є розроблення структури даних, необхідних для проведення аналізу.

Для створення інформаційної системи необхідно розробити структуру метаданих, котра описуватиме класи учасників, правила класифікації та риси поведінки учасників.

Розроблена інформаційна модель зображена на рис. 4.

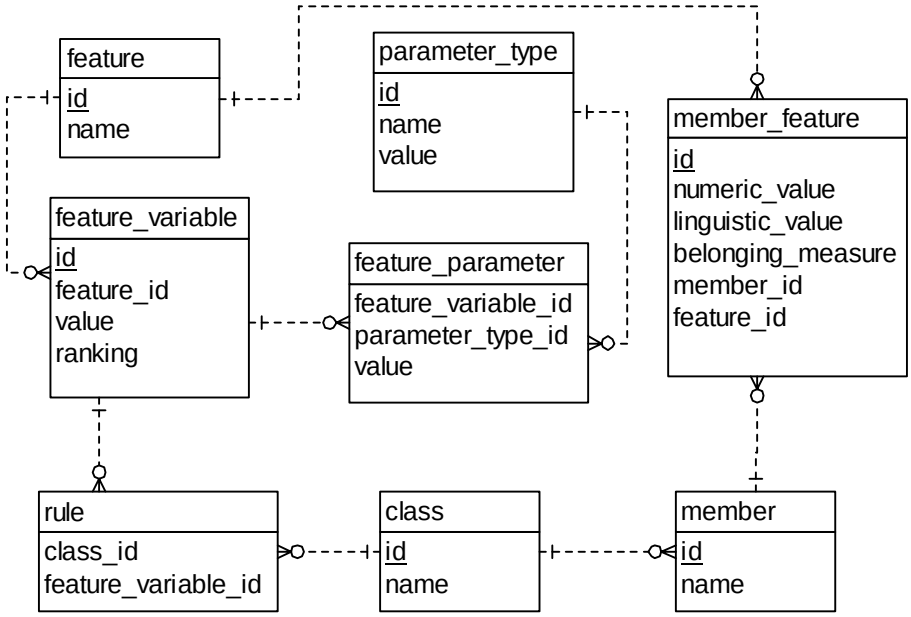

Рис. 4. Схема даних інформаційної системи аналізу поведінки учасників Веб-спільноти

Розроблена схема даних інформаційної системи аналізу поведінки учасників Веб-спільноти містить дані, необхідні для класифікації учасників Веб-форуму. Вона інтегрується з базою даних СУІН Веб-форуму на основі сутності Автор і складається з таких таблиць:

- таблиця rf_class – описує усі класи учасників Веб-спільноти;
- таблиця rf_rule – описує правила класифікації учасників Веб-спільноти;
- таблиця rf_member_feature – призначена для зберігання даних про усі риси учасників;
- таблиця rf_member_feature_variable – описує множину значень кожної лінгвістичної змінної – риси учасника Веб-спільноти.
- таблиця rf_member_feature_parameter – містить інформацію про числові межі значень кожної лінгвістичної змінної;
- таблиця rf_member_feature_parameter_type – таблиця для розрахунку рис учасників Веб-спільноти.
-

Обчислення числових характеристик учасника

Відповідно до формул, наведених у роботі [2], риси учасників Веб-форуму розраховуються так:

Активність створення дискусій усіх учасників:

```
SELECT
member_id,
```

```
count(*)*100.0/(SELECT count (*) FROM Thread) AS ThreadActivity
FROM Thread
GROUP BY member_id
```

Активність створення опитувань усіх учасників розраховутиметься так:

```
SELECT
member_id,
count(*)*100.0/(SELECT count(*) FROM Thread WHERE is_poll='y') AS PollActivity
FROM Thread
WHERE is_poll='y'
GROUP BY member_id
```

Активність створення повідомлень усіх учасників представимо як відношення усіх повідомлень учасника до загальної кількості повідомлень, створених на форумі:

```
SELECT
member_id,
count(*)*100.0/(SELECT count(*) FROM Post) AS PostActivity
FROM Post
GROUP BY member_id
```

Активність участі в опитуваннях.

```
SELECT
member_id,
count(*)*100.0/
(SELECT count(*) FROM Thread WHERE is_poll='y') as VoteActivity
FROM Vote
```

Усі види активності можна розраховувати за весь час існування спільноти, проте інколи доцільніше розраховувати активність учасників на певних часових проміжках.

Наприклад, для визначення активності створення повідомлень за роками використовуватимо такий запит:

```
SELECT
member_id,
Year(dateline) as Years,
count (*)*100.0/(SELECT count (*) FROM Post) AS PostActivityYear
FROM Post
GROUP BY member_id, Year(dateline)
```

Атрактивність розраховується двома способами.

Розрахунок атрактивності **першим** способом – середнє арифметичне відношень кількості учасників, що відреагували на появу повідомлення до загальної кількості учасників:

```
SELECT
F.member_id,
COUNT (*)*100,0/
      (SELECT COUNT (*) FROM Member)/
      (SELECT COUNT (*) FROM Post as P
       WHERE P. member_id= F.member_id)
FROM Feedback as F
GROUP BY F.member_id
```

Розрахунок атрактивності **другим** способом – як відношення кількості відповідей у дискусіях, які створив учасник, до загальної кількості відповідей у всіх дискусіях

```
SELECT
member_id,
SUM (replies)*100.0/((SELECT SUM (replies) FROM Thread)) as atractivity
FROM Thread
GROUP BY member_id
```

Креативність розраховується двома способами.

Розрахунок креативності **першим** способом: середнє арифметичне відношень кількості позитивних відгуків (feedback) до загальної кількості відгуків на кожне з повідомлень учасника:

```
SELECT
F1.member_id,
COUNT (*)*100,0/
        (SELECT COUNT (*) FROM Feedback as F2
         WHERE F2.member_id= F1.member_id)/
        (SELECT COUNT (*) FROM Post as P
         WHERE P. member_id= F1.member_id)
FROM Feedback as F1
WHERE F1.value=1
GROUP BY F1.member_id
```

Розрахунок креативності **другим** способом: відношення суми переглядів дискусій, створених учасником, до загальної суми переглядів усіх дискусій форуму:

```
SELECT
member_id,
SUM (views)*100.0/(SELECT sum (views) FROM Thread) as creativity
FROM Thread
GROUP BY member_id
```

Усі результати виконання цих запитів заносяться у таблицю member_feature, у якій зберігатимуться числові характеристики учасників (рис. 4).

Обчислення значень лінгвістичних змінних та мір приналежності

Першим етапом обчислення є визначення проміжку лінгвістичної змінної, до якого потрапляє характеристика учасника.

В інформаційній системі для визначення проміжку, до якого потрапляє числова характеристика учасника, було реалізовано функцію BorderDefinition, основою якої є такий запит:

```
SELECT
    p1.feature_variable_id,
    p1.[value] as leftvalue,
    t1.[value] as lefttype,
    p2.[value] as rightvalue,
    t2.[value] as righttype
FROM dbo.rf_member_feature_parameter as p1
INNER JOIN dbo.rf_member_feature_variable as v on v.id = p1.feature_variable_id
INNER JOIN dbo.rf_member_feature_parameter_type as t1 on p1.parameter_type_id=t1.id
INNER JOIN dbo.rf_member_feature_parameter_type as t2 on t2.[value]=t1.[value]+1
INNER JOIN dbo.rf_member_feature_parameter as p2 on p2.parameter_type_id=t2.id AND
p2.feature_variable_id=p1.feature_variable_id
WHERE
        v.memberfeature_id = @feature AND
        p1.[value] = (SELECT MAX(p2.[value])
FROM dbo.rf_member_feature_parameter as p2
WHERE p2.feature_variable_id = p1.feature_variable_id AND p2.[value] <= @card)
```

Це допоміжна функція, на основі якої визначаються нечіткі значення лінгвістичних змінних. Для визначення нечітких значень лінгвістичних змінних авторами реалізована функція LinguisticName, основою якої є два такі запити:

```
INSERT INTO @Measures
SELECT
    feature_variable_id,
```

```
        CASE
            WHEN lefttype=1 AND rightvalue<>leftvalue THEN (@card-
leftvalue)/(rightvalue-leftvalue)
            WHEN lefttype=2 OR rightvalue=leftvalue THEN 1
            WHEN lefttype=3 AND rightvalue<>leftvalue THEN (rightvalue-
@card)/(rightvalue-leftvalue)
        END as belonging_measure
    FROM borderdefinition(@feature, @card)

    SELECT TOP 1
        feature_variable_id
    FROM @Measures as m
    INNER JOIN rf_member_feature_variable as v on v.id=m.feature_variable_id
    WHERE belonging_measure = (SELECT MAX(belonging_measure) FROM @Measures)
    ORDER BY ranking DESC
```

Наступним кроком є визначення ступеня приналежності характеристики учасника до значення лінгвістичної змінної.

Для цього була розроблена функція BelongingMeasure, котра ґрунтується на таких двох запитах:

```
    INSERT INTO @Measures
        SELECT
            feature_variable_id,
            CASE
                WHEN lefttype=1 AND rightvalue<>leftvalue THEN (@card-
leftvalue)/(rightvalue-leftvalue)
                WHEN lefttype=2 OR rightvalue=leftvalue THEN 1
                WHEN lefttype=3 AND rightvalue<>leftvalue THEN (rightvalue-
@card)/(rightvalue-leftvalue)
            END as belonging_measure
        FROM borderdefinition(@feature, @card)
```

та

```
    SELECT MAX(belonging_measure) FROM @Measures
```

Якщо характеристика учасника потрапляє у проміжок, який належить до двох значень лінгвістичної змінної, то учаснику буде присвоєне те, ступінь приналежності якого більший.

Значення лінгвістичних змінних та мір приналежності для усіх рис усіх учасників Веб-спільноти обчислюється за допомогою такого запиту:

```
    SELECT
    member_id,
    feature_id,
    linguisticname(feature_id,numeric_value),
    belongingmeasure (feature_id, numeric_value)
    FROM member_feature
```

Результати виконання запиту заносяться в таблицю member_feature.

Наступним кроком є класифікація учасників на основі визначених рис відповідно до правил, нведених у [2].

Класифікація учасників в інформаційній системі аналізу поведінки учасників Веб-спільнот реалізована за допомогою функції Classify на основі правил, котрі зберігаються в таблицях (rf_class та rf_rule).

Приклади реалізації

Система розроблена засобами СУБД Microsoft SQL Server. Апробація системи проводилася на основі Веб-спільноти Форуму Рідного Міста (http://misto.ridne.net/), який функціонує на основі СУІН ХМВ.

Продемонструємо роботу системи, вибравши для аналізу чотирьох випадкових учасників з кодами з проміжку 5–12. У цьому проміжку є чотири учасники з такими кодами: № 8, № 9, № 10, № 11.

Функція приналежності нечітких множин характеристик учасників виглядає так:

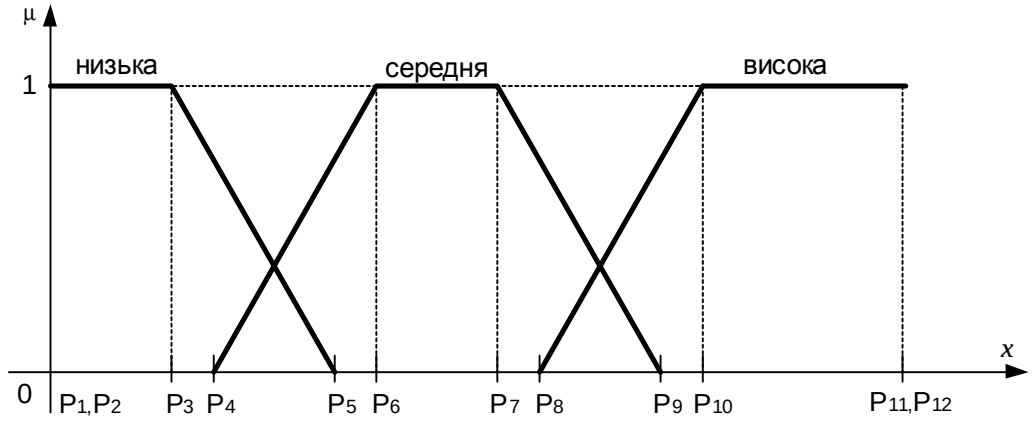

Рис. 5. Функція приналежності нечітких множин

Таблиця параметрів для визначення нечітких значень характеристик виглядає так:

Функція приналежності	Точки проміжку	Точка	Величина
Низька	Нижній поріг	P_1	0
	Перший пік	P_2	0
	Другий пік	P_3	0,1
	Верхній поріг	P_5	0,22
Середня	Нижній поріг	P_4	0,16
	Перший пік	P_6	0,35
	Другий пік	P_7	0,6
	Верхній поріг	P_9	0,8
Висока	Нижній поріг	P_8	0,7
	Перший пік	P_{10}	0,9
	Другий пік	P_{11}	100
	Верхній поріг	P_{12}	100

Першим кроком був розрахунок числових величин характеристик учасника. Результати виконання запитів занесені в таблицю member_feature, частина якої виглядає так:

Таблиця 1

Розрахунок числових величин характеристик учасника

Код учасника	Характеристика	Величина
1	2	3
8	активність ств. повідомлень	0.897450234846
8	активність ств. дискусій	1.124818577648
8	атрактивність	0.945275867751
8	креативність	1.249002807237

Продовження табл. 1

1	2	3
9	активність ств. повідомлень	0.422165063744
9	активність ств. дискусій	0.997822931785
9	атрактивність	0.585741464803
9	креативність	0.972568791464
10	активність ств. повідомлень	0.114627600089
10	активність ств. дискусій	0.689404934687
10	атрактивність	0.239689601965
10	креативність	0.501121249686
11	активність ств. повідомлень	0.166349809885
11	активність ств. дискусій	0.108853410740
11	атрактивність	0.281635282309
11	креативність	0.253319759087

Для зручності представлення числові величини характеристик розраховується у відсотках.

Наступними кроками є обчислення значення лінгвістичної змінної і ступеня приналежності.

Таблиця 2

Розрахунок значення лінгвістичної змінної і міри приналежності

Код учасника	Характеристика	Величина	Значення лінгвістичної змінної	Міра приналежності
8	Креативність	1.249002807237	висока	1.000
8	Активн. ств. диск.	1.124818577648	висока	1.000
8	Атрактивність	0.945275867751	висока	1.000
8	Активн. ств. пов.	0.897450234846	висока	0.985
9	Активн. ств. диск.	0.997822931785	висока	1.000
9	Креативність	0.972568791464	висока	1.000
9	Атрактивність	0.585741464803	середня	1.000
9	Активн. ств. пов.	0.422165063744	середня	1.000
10	Активн. ств. диск.	0.689404934687	середня	0.555
10	Креативність	0.901121249686	висока	1.000
10	Атрактивність	0.739689601965	висока	1.000
10	Активн. ств. пов.	0.114627600089	низька	0.875
11	Атрактивність	0.281635282309	середня	0.642
11	Креативність	0.253319759087	низька	1.000
11	Активн. ств. пов.	0.166349809885	низька	0.450
11	Активн. ств. диск.	0.108853410740	низька	0.925

Останнім етапом є класифікація учасників Веб-форуму. Провівши класифікацію, ми визначаємо, чи належить учасник хоча б до одного з класів.

Результати класифікації виглядають так:

Таблиця 3

Класифікація учасників

Код учасника	Клас
8	Активіст
9	Активіст
10	Автор
11	Читач

Отже, як бачимо з наведених результатів, два учасники належать до класу «Активіст» (володіють високим рівнем активності і атрактивності), один учасник належить до класу «Автор» (володіє високим рівнем креативності і атрактивності). Ще один учасник належить до класу «Читач» (має низьку активність і креативність).

Висновки

Створення інформаційної системи аналізу поведінки учасників спільноти є важливим і актуальним завданням, оскільки Веб-спільноти є дуже популярним явищем, а управління ними – дуже складним завданням.

Проаналізовано структури наявних СУІН Веб-форумів, досліджено процес аналізу поведінки учасників Веб-спільнот, побудовано схему бази даних інформаційної системи аналізу поведінки учасників Веб-спільноти, здійснено програмну реалізацію алгоритмів аналізу поведінки учасників Веб-форумів.

Система, розроблена засобами СУБД Microsoft SQL Server, дає змогу власникам та адміністраторам Веб-форумів аналізувати поведінку членів спільноти та приймати правильні рішення в процесі модерування спільноти, тобто зробити управління спільнотою простішим та ефективнішим.

1. Kravets. R. Typical ways of web-communities development / R. Kravets, A.M. Peleschyshyn, Yu. Syerov // Proceedings of the International Conference on Computer Science and Information Technologies, CSIT`2006, September 28th-30th, Lviv, Ukraine, p. 56–58. 2. Сєров Ю.О. Моделювання поведінки та класифікація учасників Веб-спільнот на основі нечітких множин // Вісник Нац. ун-ту «Львівська політехніка». – 2008. – № 610 – С. 218–228. 3. Peleschyshyn A. Typical ways of web-communities development / A. Peleschyshyn, Yu. Syerov // Proceedings of the International Conference on Computer Science and Information Technologies, CSIT`2006, September 28th-30th, Lviv, Ukraine, p.56–58. 4. Пасічник В.В. Аналіз критеріїв ефективності Веб-спільнот на основі форумів / В.В. Пасічник, Р.Б. Кравець, Ю.О. Сєров // Материалы 12-го Международного молодежного форума «Радиоэлектроника и молодежь в XXI веке», 1-3 апреля 2008 г. Ч. 2, с.493, Харьков. 5. R. Kravets Information System of Web Community Members Behavior Analyzing and Classifying / R. Kravets, Yu. Syerov// Proceedings of the international conference on computer science and information technologies (CSIT' 2008), September 25th-27th, Ukraine, Lviv, 2008. – p. 41–44. 6. Круглов В.В. Нечеткая логика и искусственные нейронные сети: Учеб. пособие / В.В. Круглов, М.И. Дли, Р.Ю. Голунов. – М.: Изд-во Физ.-мат. лит-ры, 2001. – 224 с.