УДК 513.7

Т. Г. Ємел'яненко, А. О. Домашич

Дніпропетровський національний університет імені Олеся Гончара

ДОСЛІДЖЕННЯ МОДЕЛЕЙ СОЦІАЛЬНИХ МЕРЕЖ

Досліджено моделі соціальної рекомендації новин, які поєднують у собі минулу діяльність користувачів і соціальні відносини для отримання рекомендацій. Запропоновано можливі шляхи покращення результатів гібридних методів та надано рекомендації щодо специфіки соціальної мережі.

Ключові слова: *соціальна мережа; соціальна рекомендація; модель соціальної мережі; інформаційна технологія.*

Исследованы модели социальной рекомендации новостей, объединяющие прошлую деятельность пользователей и социальные отношения для получения рекомендаций.

Ключевые слова: *социальная сеть; социальная рекомендация; модель социальной сети; информационная технология.*

The models of the social news recommendation were created. It was used combination of past user activities and social relations to receive recommendations.

Keywords: *social network; social recommendation; social network model; information technology.*

Вступ. Соціальні мережі у житті сучасного суспільства відіграють не останню роль у формуванні відносин між людьми і поширенні інформації. Пошук контенту, що відповідає інтересам користувачів, можливо здійснювати за допомогою рекомендаційних систем, вони прогнозують вподобання користувачів для впровадження персоніфікованих рекомендацій. Соціальна рекомендація стає новим підходом, який безпосередньо використовує соціальні зв'язки між членами суспільства для виявлення нових цікавих джерел інформації.

Аналіз останніх досліджень і публікацій. Рекомендаційні методи можна розділити на три основні категорії: на основі змісту (content-based), спільної фільтрації (collaborative filtering) і соціальні методи фільтрації (social filtering). У [1] розглянуто алгоритм user-based collaborative filtering – розширена версія алгоритму collaborative

filtering . Його основна ідея базується на обчисленні міри схожості між парами користувачів. У [2] представлено потужний механізм для collaborative filtering та обробки даних користувачів, метод досягає конкурентоспроможної рекомендації та точності прогнозування, є високомасштабованим та гнучким. Декомпозиція вподобань користувачів з використанням груп користувачів, що перекриваються, є відмінною рисою цього підходу у порівнянні з традиційними методами, які базуються на запам'ятовуванні та моделюванні. У [3] висувається припущення, що минула діяльність користувачів відіграє менш важливу роль, ніж соціальний вплив: рекомендації, отримані абстрактним математичним аналізом, люди цінують менше ніж ті, що отримали від своїх друзів або колег.

Основними характеристиками соціальних мереж є *link reciprocity* (тенденція пар користувачів утворювати зв'язки один з одним, обчислюється як відношення кількості двонапрямлених зв'язків до загальної кількості зв'язків) та *clustering coefficient* (виражає тенденцію мережі до утворення міцно зв'язаних компонентів, визначається як відношення кількості спрямованих трикутників зв'язків між користувачем та його сусідами першої черги до загальної кількості таких трикутників, що можуть бути сформованими). Соціальні мережі характеризуються великими значеннями цих показників порівняно з розподілом зв'язків у випадкових графах, що є від двох до п'яти порядків менші.

Постановка задачі. Дослідити різні стратегії вибору нових лідерів. Порівняти отримані моделі інформаційних мереж з точки зору задоволення користувачів, адаптації мережі та ефективності рекомендацій.

Основний матеріал. Розглянемо модель соціальної мережі, яка використовується для дослідження різних стратегій вибору лідерів.

Мережа складається з U користувачів, кожен з яких безпосередньо пов'язаний з L іншими користувачами (його лідерами). Значення L є фіксованим, бо зазвичай користувачі стежать за обмеженою кількістю джерел. Користувачі отримують новини від своїх лідерів і з часом читають і оцінюють їх. Крім того, вони можуть самостійно додавати новий контент у систему.

Оцінка новини α користувачем i ($e_{i\alpha}$) може бути +1 (сподобалося), -1 (не сподобалося), або 0 (ще не читав). Схожість читацьких смаків користувачів i та j (s_{ij}) оцінюється шляхом порівняння минулих оцінок користувачів. Якщо i та j разом оцінили

N_{ij} новин, і погодилися A_{ij} раз, подібність їх смаків може бути виміряна з точки зору ймовірності досягнення згоди взагалі.

$$N_{ij} = \frac{A_{ij}}{N_{ij}}\left(1 - \frac{1}{\sqrt{N_{ij}}}\right), \quad (1)$$

де множник у дужках зменшує вплив пар користувачів з малим перекриттям N_{ij}. Для $N_{ij} \le 1$, s_{ij} замінюється невеликим додатним значенням s_0.

Поширення новин відбувається таким чином. Коли в момент t_α користувач i додає новину α в систему, вона передається до його j читачів з рекомендаційним показником, пропорційним схожості читацьких смаків s_{ij}. Якщо пізніше ця новина сподобалася одному з читачів, вона так само передається далі до k читачів j користувача (з рекомендаційним показником пропорційним s_{jk}) і так далі. У момент часу t для будь-якого користувача k новина α є рекомендованою відповідно до її поточного рекомендаційного показника [4]:

$$R_{k\alpha}(t) = \delta_{e_{k\alpha},0}\lambda^{t-t_\alpha}\sum_{l\in L_k} s_{kl}\delta_{e_{l\alpha},1}, \quad (2)$$

де L_k – множина лідерів користувача k. Вираз $\delta_{e_{k\alpha},0}$ дорівнює 1 лише тоді, коли користувач k ще не читав новину α. Вираз $\delta_{e_{l\alpha},1}$ дорівнює 1 тільки якщо користувачу l сподобалася новина α. Сума описує випадок, коли користувач отримує одну й ту саму новину від кількох лідерів одночасно. У цьому випадку рекомендаційні показники підсумовуються, віддзеркалюючи той факт, що новина, яка сподобалася кільком лідерам, швидше за все, сподобається і цьому користувачеві також. Нарешті, щоб зробити свіжі новини швидкодоступними, рекомендаційні показники експоненційно згасають з часом, де $\lambda \in (0;1]$ – коефіцієнт загасання.

Після вибору початкової випадкової конфігурації мережі (випадковий вибір лідерів) зв'язки періодично оновлюються, спрямовуючи систему до оптимального стану, де користувачі з високою схожістю смаків безпосередньо пов'язані між собою. Коли відбувається оновлення зв'язків, для i користувача лідер з найнижчим значенням схожості j замінюється на нового користувача k, якщо

$s_{ik} > s_{ij}$. Існують різні стратегії вибору нових лідерів.

Випадкове оновлення зв'язків. Користувач k є випадково обраним у мережі.

Локальне оновлення зв'язків. Користувач k обирається як користувач з максимальним значенням s_{ik} в околі користувача i . Цей механізм базується на спостереженні, що два користувачі, які мають спільних знайомих, швидше за все, мають схожі інтереси. Існують різні способи визначення такого околу.

Гібридне оновлення зв'язків. Поряд з локальним оновленням зв'язків у деяких випадках використовується і випадкове оновлення. Цей механізм імітує пошук друзів серед друзів друзів (локальне оновлення), а також можливість випадкових зустрічей, які можуть призвести до довгострокових відносин (випадкове оновлення).

Глобальне оновлення зв'язків. k – це користувач, що максимізує s_{ik} серед усіх U користувачів (це локальне оновлення зв'язків, де околом вважається вся мережа).

У моделі, що описана вище, процедура оновлення лідерів призначена для спрямування мережі до оптимального стану, де користувачі з високою схожістю читацьких смаків безпосередньо пов'язані між собою. Таким чином, система здатна ефективно рекомендувати саме цікаві новини відповідним користувачам. Треба нагадати, що модель мережі обмежена правилами локального пошуку, тому що глобальний механізм пошуку (глобальне оновлення зв'язків) є дуже вимогливим для масштабних мереж, а також є неможливим без централізованого контролю. З іншого боку, стратегія випадкового оновлення зв'язків є дуже неефективною, тому що нові кращі лідери навряд чи можуть бути знайдені випадково. Крім того, у реальних умовах користувачі мають доступ лише до обмеженої частини мережі. Тому виникає необхідність визначити поняття «околу» користувача, тобто підмножини близьких користувачів, які є кандидатами в лідери. Це основа локального оновлення зв'язків. Вибір конкретного околу повинен бути достатньо ефективним, щоб дозволяти користувачам знайти своїх однодумців. Наприклад, набір кандидатів у лідери не повинен бути занадто великий, так як у цьому випадку пошук стає некерованим як з боку системи, так і з боку користувача. З іншого боку, якщо розмір околу дуже малий (у порівнянні з цілою мережею), оновлення зв'язків може зупинитися в локальному оптимумі: еволюція топології зупиняється, якщо найкращі лідери користувачів у певний момент часу знаходяться за межами їх околів (вони не можуть бути

досягнуті). Це означає, що алгоритм потрапив у пастку в неоптимальному стані. Можливим рішенням цієї проблеми є використання деякого відсотка випадковості у виборі, як у гібридному оновленні зв'язків. Таким чином, користувачі можуть встановити зв'язок незалежно від відстані між ними, а набір потенційних кандидатів у лідери для кожного користувача є цілою мережею. У подальшому аналізі головна увага буде приділятися саме гібридному оновленню зв'язків із 10 % випадковістю. Це дозволить детально розглянути локальний пошук та уникнути зупинки у локальному мінімумі.

Смаки користувача i представлені D -вимірним бінарним вектором t_i . Атрибути новини α представлені D -вимірним бінарним вектором w_α . Кожен вектор має фіксовану кількість D_W елементів, що дорівнюють одиниці (активні смаки), а всі інші елементи дорівнюють нулю. Користувач з унікальним набором активних смаків представлений у мережі лише один раз, тобто $U = \begin{pmatrix} D \\ D_W \end{pmatrix}$. Це також означає, що вектори смаків двох користувачів відрізняються принаймні у двох елементах. Оцінка новини α користувача i ґрунтується на перекритті вектора смаків користувача з вектором атрибутів новини

$$\Omega_{i\alpha} = (t_i, w_\alpha), \tag{3}$$

де $(\cdot, \cdot)$ — скалярний добуток двох векторів. Якщо $\Omega_{i\alpha} \geq \Delta$, користувачеві i подобається новина α ($e_{iw} = 1$), в іншому випадку – не подобається ($e_{iw} = -1$). Тут Δ – поріг симпатій користувачів.

Моделювання виконується в дискретному часі. На кожному кроці окремий користувач є активним з ймовірністю p_W. Якщо він активний, користувач читає і оцінює перші R новин зі свого списку рекомендацій, та з імовірністю p_S публікує власні новини з атрибутами, ідентичними активним смакам користувача. Оновлення зв'язків відбувається кожні u кроків. Поведінку системи можна описати таким алгоритмом.

Алгоритм 1 [4].

1. Ініціалізація. Заповнити мережу користувачами, вектори смаків яких є різноманітними розміщеннями з D елементів по D_W. Тобто розмір мережі можна обчислити за формулою

$$C_D^{D_W} = \frac{D!}{D_W!(D - D_W)}. \quad (4)$$

Кожного користувача мережі i пов'язати з L випадковими користувачами (призначити лідерів).

2. Крок моделювання. Для кожного користувача мережі i: якщо $random(0;1) \le p_W$, то i користувач є активним; якщо i користувач активний, то оцінити перші R новин зі списку рекомендацій за формулою (1); якщо $random(0;1) \le p_S$, то i користувач публікує новину.

3. Оновлення зв'язків. Для кожного користувача мережі i: обрати лідера з найменшою величиною схожості смаків; серед сусідів i-го користувача знайти лідера з найбільшою схожістю смаків (тут під сусідами розуміється група користувачів, визначена типом околу); якщо новий лідер є ближчим за читацькими вподобаннями, замінити найгіршого з лідерів на нового.

Крок 2 повторюється задану кількість разів. Крок 3 повторюється кожні u кроків. Статистичні дані також фіксуються кожні u кроків.

Для оцінки продуктивності системи використано такі показники:

- LR (link reciprocity) – тенденція пар користувачів до утворення взаємних зв'язків, тобто співвідношення кількості двонапрямлених зв'язків до загальної кількості зв'язків у мережі.

$$LR = \frac{BL}{OL}, \quad (5)$$

де BL (bidirectional links) – кількість двонапрямленних зв'язків, OL (onedirectional links) – загальна кількість зв'язків у мережі.

- AF (approval fraction) – співвідношення новин, що сподобалися користувачеві, до загальної кількості рекомендованих новин (6). Цей показник віддзеркалює, як часто користувачі задоволені новинами, які вони читають

$$AF = \frac{NA}{NE}, \quad (6)$$

де NA (news approved) – кількість новин, які користувач оцінив позитивно, NE (news evaluated) – загальна кількість рекомендованих новин.

- AD (average differences) – середнє число елементів векторів смаків, у яких користувачі відрізняються від своїх лідерів (7). Цей показник вимірює, наскільки добре мережа адаптована до смаків користувачів.

$$AD = \sum_{j=1}^{L} \sum_{i=1}^{D} d_{ij}\ , \qquad (7)$$

le $d_{ij} = \begin{cases} 1, D_i^U = D_i^{L_j} \\ 0, D_i^U \neq D_i^{L_j} \end{cases}, D_i^U$ – елемент вектору смаків користувача, $D_i^{L_j}$ – елемент вектора смаків j-го лідера користувача, d_{ij} – різниця у векторі смаків користувача та його j-го лідера.

Під час моделювання поведінки мережі було використано такі значення параметрів:

- кількість лідерів кожного користувача мережі $L = 10$;
- імовірність активності користувача $p_W = 0{,}05$;
- імовірність публікації новини $p_S = 0{,}02$;
- кількість новин, що оцінює користувач на кожному кроці моделювання $R = 3$;
- коефіцієнт загасання $\lambda = 0{,}9$;
- мінімальна схожість інформаційних уподобань користувачів $s_0 = 0{,}001$;
- збір статистики кожні $u = 10$ кроків.

Було проведено дві групи експериментів з параметрами. Моделювання № 1: розмір вектора інтересів користувачів $D = 12$, кількість активних інтересів $D_W = 5$, кількість користувачів у мережі $U_1 = 792$, порогове значення при оцінці новин $\Delta = 2$. Моделювання № 2: $D = 14,\ D_W = 6,\ U_2 = 3003,\ \Delta = 3$.

Головну увагу приділено гібридним методам пошуку нових лідерів. Результати моделювання з використанням глобального та випадкового пошуку наведено в якості граничних випадків. В обох експериментах мережі, що реалізують методи гібридного і глобального пошуку, за показником link reciprocity (LR) можна розглядати як соціальні. Значення показника швидко досягають рівня, що є значно більшим за значення аналогічного показника у графі з випадковим розподіленням зв'язків, тобто у мережі з випадковим вибором нових лідерів. У початковій структурі мережі, де зв'язки встановлюються випадковим чином, імовірність утворення двонаправленого зв'язку дорівнює звичайній математичній імовірності утворення зв'язку між двома

користувачами, тобто: $LR^0 = \frac{L}{U-1}$. Для проведених експериментів $LR_1^0 = 0{,}0126$ та $LR_2^0 = 0{,}0058$ відповідно. З плином часу та реогранізацією структури мережі показник LR зростає. Отже, після досягнення показником LR цих граничних значень мережу можна вважати соціальною (рис. 1, 2).

.

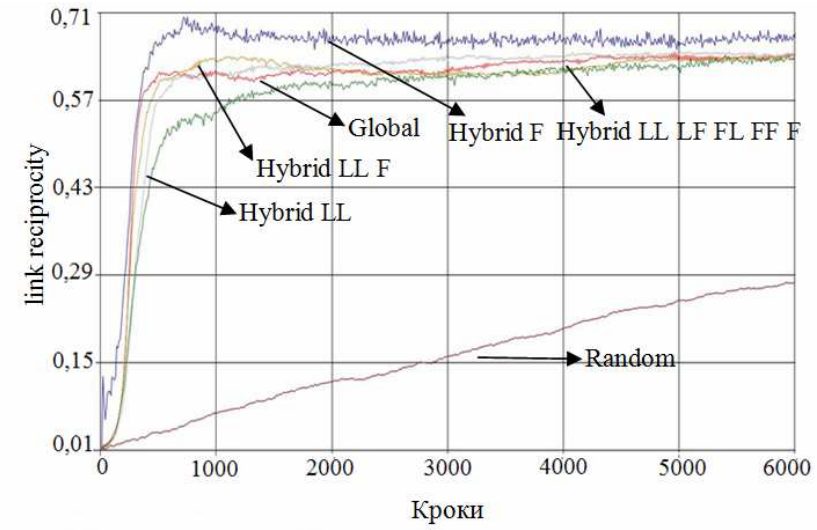

Рисунок 1 – Графік показника LR_1 залежно від часу $(U_1 = 792)$

Як і очікувалося, показник LR є найбільшим у мережі, що реалізує вибір нових лідерів у множині читачів користувача, тобто F. Цей показник має найменше значення у мережі з випадковим вибором нових лідерів. Інші методи досягають схожих значень LR, що є порівняним з кількістю взаємних зв'язків у реальних соціальних мережах.

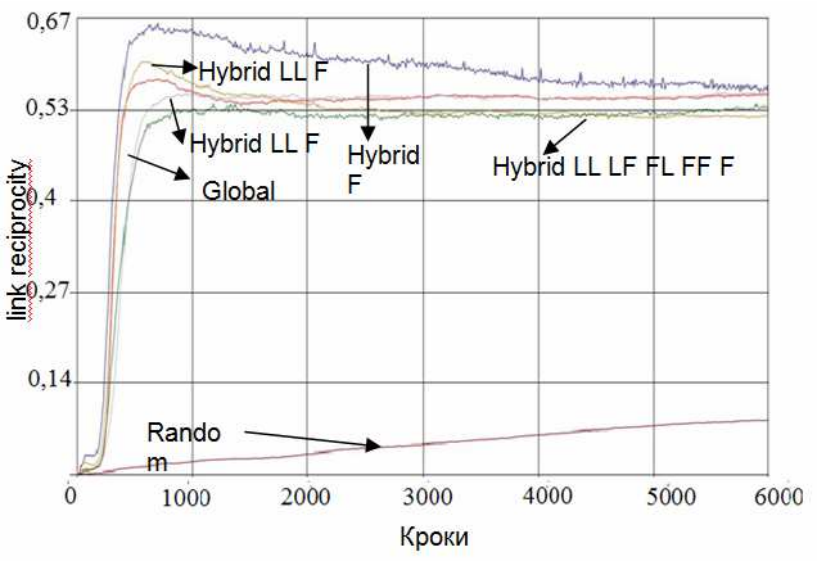

Рисунок 2 – Графік показника LR_2 залежно від часу $(U_2 = 3003)$

Розглянемо більш детально показник AF (approval fraction), тобто якість рекомендації новин. Очевидно, що при початковій випадковій конфігурації мережі, на перших кроках моделювання відбуваються значні зміни у складі лідерів, а отже і якості новин, що потрапляють до рекомендаційного списку користувачів. Враховуючи, що початкові значення AF знаходяться під впливом випадкових факторів, їх можна виключити з розгляду. Як бачимо, показник AF має більше значення при меншому розмірі мережі (рис. 3, 4). 92–95 % новин позитивно оцінюються їх читачами для мережі з 792 користувачами, і лише 86–89 % – для мережі з кількістю 3003 особи. Це має природне пояснення: у першому експерименті задача рекомендації цікавого контенту спрощується завдяки меншому розміру мережі.

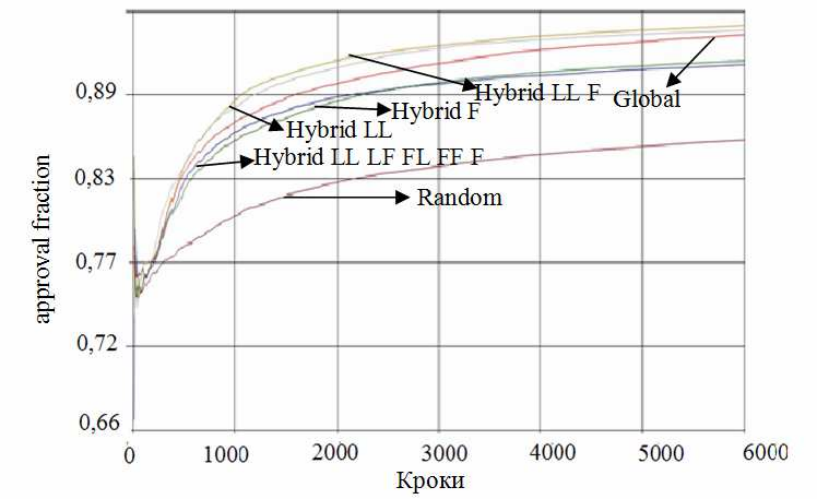

Рисунок 3 – Графік показника AF_1 залежно від часу $(U_1 = 792)$

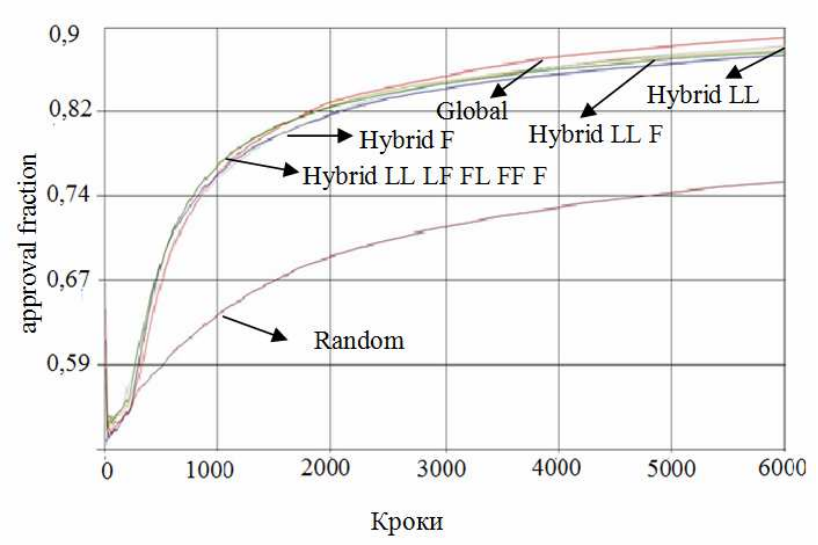

Рисунок 4 – Графік показника AF_2 залежно від часу $(U_2 = 3003)$

Щодо порівняння якості різних стратегій гібридного пошуку, невеликий розмір моделей соціальних мереж та наявність тенденції до зменшення різниці в якості методів не дають змогу виокремити найкращий метод гібридного пошуку. Головною метою даного дослідження є показати, що в умовах, де реалізація глобального методу пошуку нового лідера є неможливою з практичної точки зору через вимогливість до обчислювальних потужностей серверів та часу, методи гібридного пошуку дають високі результати і можуть бути застосовані в якості методів соціальної рекомендації. Під час аналізу зміни параметра AD (average differences) виявлено, що найшвидше найкращих показників досягає метод глобального пошуку (рис. 5, 6). Цей результат є очевидним, але недосяжним у реальних умовах. Методи гібридного локального пошуку дають близькі за якістю результати, що підтверджує висновки, зроблені вище.

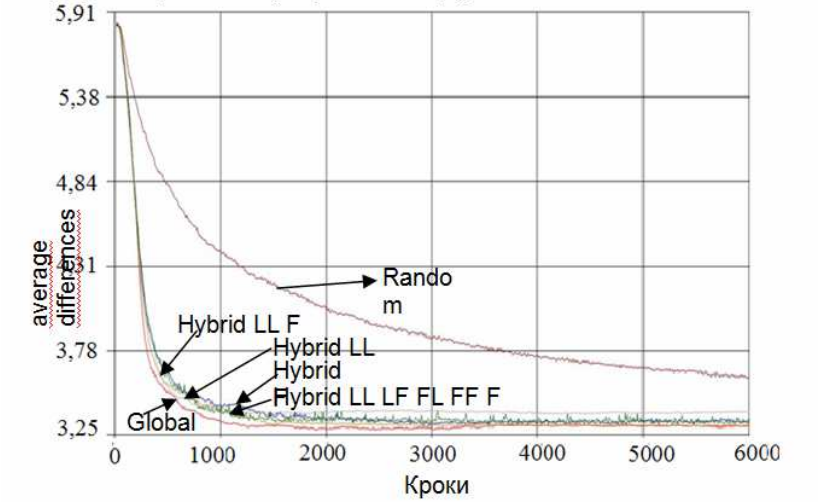

Рисунок 5 – Графік показника AD_1 залежно від часу $(U_1 = 792)$

Проведені експерименти свідчать про те, що побудована модель відповідає реальній структурі соціальних мереж. Було розглянуто результати декількох стратегій гібридного локального пошуку, проте важко виокремити один метод, який дає абсолютно найкращі показники.

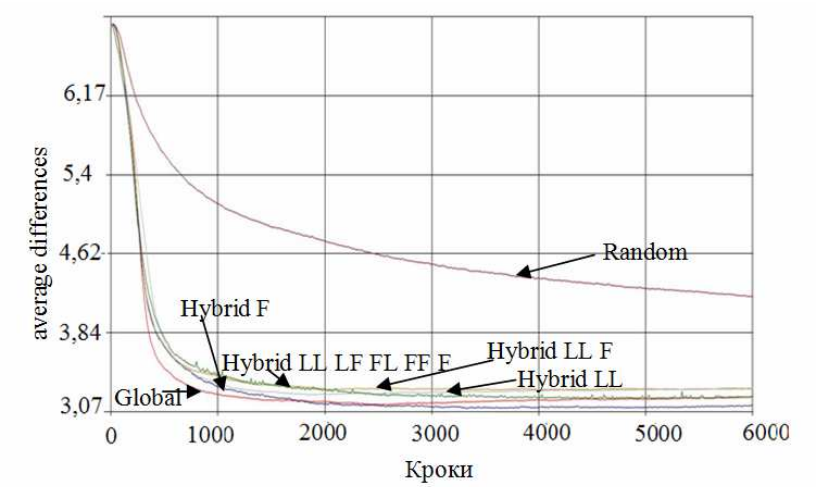

Рисунок 6 – Графік показника AD_2 залежно від часу $(U_2 = 3003)$

Під час вибору конкретної стратегії локального пошуку слід, крім результатів моделювання, враховувати специфіку соціальної мережі. Наприклад, у соціальній мережі linkedin.com, що спрямована на професійну діяльність користувачів, кожен прагне зв'язатися з якнайбільшою кількістю друзів, щоб привернути до себе увагу та, можливо, отримати цікаві пропозиції щодо місця роботи. У такій мережі доцільним буде використання гібридного методу LL+LF+FL+FF+F, що включає до набору потенційних лідерів якомога більше користувачів. І навпаки, у таких мережах як pinterest.com, тобто присвячених конкретним захопленням користувачів, кращою стратегією представляється LL+F, де кожен користувач цікавиться лише обмеженим колом рекомендацій, які найвірогідніше йому сподобаються.

Висновки. Розроблено інформаційну технологію соціальної рекомендації новин, яка поєднує в собі минулу діяльність користувачів і соціальні відносини. Вона імітує процес поширення новин, характерний для соціальних систем, в яких структура зв'язків постійно еволюціонує з плином часу. Еволюція топології поліпшує задачу користувачам, які шукають нові й кращі джерела інформації. Оскільки глобальна оптимізація зв'язків є обчислювально складною для великої системи, ключовим питанням є знаходження хороших нових лідерів для користувачів. Розглянуто методи випадкового, локального,

гібридного та глобального оновлення зв'язків. Наведений алгоритм описує поведінку системи у часі. Запропоновані автоматизовані абстрактні правила допомагають створювати мережі, які не тільки є ефективними для поширення інформації, але і за структурою нагадують результати реальної людської діяльності. Було побудовано декілька моделей, що реалізують різні стратегії гібридного локального пошуку, де кожен користувач має обмежене уявлення про мережу, та за параметрами є наближеними до реальних соціальних мереж. Проведені експерименти свідчать про те, що побудована модель відповідає реальній структурі соціальних мереж. Отримані результати свідчать, що в умовах обмеженості ресурсів машинного часу, коли реалізація методу глобального пошуку не є можливою, методи гібридного локального пошуку з різними стратегіями вибору лідера дають високі результати. Запропоновано можливі шляхи покращення результатів гібридних методів та надано рекомендації щодо специфіки соціальної мережі.

Бібліографічні посилання

1. **Yamashita A., Kawamura H., Suzuki K.** Similarity Computation Method for Collaborative Filtering Based on Optimization // Journal of Advanced Computational Intelligence and Intelligent Informatics, 2010. Vol. 14. № 6. P. 654–660.

2. **Hoffman T.** Component-based LDA Face Descriptor for Image Retrieval // ACM Transactions on Information Systems, 2004. Vol. 22. № 1. P. 89–115.

3. **Cimini G., Medo M., Zhou T., Wei D., Zhang Y.-C.** Heterogeneity, quality, and reputation in an adaptive recommendation model // The European Physical Journal B. 2011. Vol. 80. № 2. P. 201–208.

4. **Chen D., Giulio C., Linyuan L., Medo M., Zhang Y.-C., Zhou T.** Enhancing topology adaptation in information-sharing social networks // Physical Review E. 2012. Vol. 85. № 4.

Надійшла до редколегії 12.07.17