УДК 681.324

В. М. Дубовой, д-р. техн. наук, проф;
О. М. Москвін, асп.

МОДЕЛЬ ПОВЕДІНКИ КОРИСТУВАЧА У ГІПЕРТЕКСТІ

Запропоновано модель поведінки користувачів у гіпертекстових мережах для визначення їх маршрутів, що характеризуються оптимальними показниками протяжності та інформаційної цінності. Необхідність розробки зазначеної моделі мотивована відсутністю у відкритому друку відповідних спеціалізованих моделей. Розроблену алгебру можна застосовувати в системах оцінювання гіпертекстових мереж та для оптимізації структури їх посилань.

Вступ

Гіпертекстові системи відносяться до складних систем з точки зору моделювання. Синтаксичний і семантичний аналіз дозволяють встановити зв'язки, виділити предметні області, але вони не здатні встановити цілі з якими користувачі використовують ті або інші інформаційні сегменти гіпертекстової мережі.

Цілі користувачів щодо інформації є різними. Для деяких користувачів певний сегмент може бути лише транзитним на шляху перегляду, для інших — метою пошуку інформації. Формалізація поведінки користувачів на основі їх цілей є предметом дослідження цієї статті.

Проведений аналіз літературних джерел [1, 2], показав, що на сьогодні не існує відкритих спеціалізованих моделей, що дозволяють моделювати поведінку користувачів у гіпертекстових мережах і відповідно визначати їх цілі. Для розв'язання цієї задачі з урахуванням того, що об'єктом дослідження є складна, багатоелементна і неоднорідна динамічна система, на погляд авторів доцільно використати ентропійний підхід [3] за аналогією з міськими системами.

Метою статті є побудова математичної моделі поведінки користувачів у гіпертекстових мережах для ідентифікації маршрутів із оптимальними значеннями вартості інформаційного перегляду. Для цього пропонується використати ентропійний підхід, оскільки авторами передбачається, що користувачі складають невід'ємну складову моделі гіпертекстового інформаційного середовища і визначають важливість тих чи інших інформаційних статей та їх маршрутів, тим самим опосередковано впливаючи на посилальну організацію веб-ресурсів.

Модель поведінки користувача у гіпертексті

Гіпертекстові системи можна віднести до класу систем, в яких детермінований характер процесів пошуку інформації поєднується з їх стохастичною природою. Формальну модель таких систем будемо називати макросистемою, що описує перетворення випадкових міжелементних взаємодій в деякий регулярний процес. Цим самим в системі можна розглядати 2 рівня: мікрорівень, де вибір інформаційних статей користувачем випадковий, і макрорівень, на якому вибір інформаційних статей з точки зору інформаційних доменів є детермінованим.

Нехай задано систему з фіксованим розподілом інформаційних доменів між якими виникає деякий потік користувачів. Для опису даної системи вкажемо для кожного користувача початкову та кінцеву точку на шляху перегляду.

Розглянемо макропідхід до опису стану даної системи. Для цього, звернемося до таблиці зображеної на рис. 1. По вертикалі даної таблиці розміщено точки входу користувачів $p_{e_i}\ (i = 1,\ldots,n)$, а по горизонталі — точки виходу $p_{o_j}\ (j = 1,\ldots,n)$. Точкою входу користувача є перша сторінка перегляду, що входить у задану підмережу сторінок, що розглядається. Точкою виходу — остання сторінка переглянута користувачем з заданої підмережі сторінок.

Будемо вважати, що усі користувачі в системі є пронумерованими. Тоді станом системи є розміщення помічених користувачів у клітинах таблиці на рис. 1. При визначенні потоку між інформаційними доменами *i* та *j* важливим є лише загальна кількість T_{ij} користувачів, що здійснюють навігацію, а не їх структурний склад. Розподілом в даній системі будемо називати $T = \left\{T_{ij},\ i = 1,\ldots,n;\ j = 1,\ldots,m\right\}$. Розподіл T характеризує макровластивості системи.

Таким чином, кожна комірка матриці «входи–виходи» на рис. 1 складається з двох частин: по-перше, із множини помічених користувачів, що переміщуються з *i* в *j,* а по-друге із загальної кількості T_{ij} користувачів, що переміщуються з *i* в *j*. Існує багато станів, що призводять до одного і того ж розподілу. Якщо кожен стан системи реалізується з рівною імовірністю, тоді можна знайти найімовірніший розподіл обчисливши множину значень T_{ij}, з якими пов'язана найбільша кількість станів.

Рис. 1. Таблиця, що зв'язує точки входу з точками виходу

Будемо розглядати гіпертекстову систему як систему з максимальною корисністю. Нехай $x_1, x_2, \ldots, x_n$ — сторінки $1, \ldots, N$ у гіпертексті, за якими користувач здійснює перегляд гіпертексту з інформаційними витратами перегляду $p_1,\ p_2, \ldots, p_n$, виходячи із загального обмеження на витрати I, x_i — кількість переходів на сторінку i у маршруті.

$$X = \left(\left(x_k,\ x_n \right) \middle| x_k \in P,\ x_n \in P, k = 1, \ldots, N, n = 1, \ldots, N, k \neq n \right).$$

Визначимо вартість перегляду інформаційної статті як

$$p_i = \frac{1}{t_{p_i} \ln\left(ROC_{p_i} + 1 \right) + 1}, \tag{1}$$

де t_{p_i} — фактичний час перегляду сторінки; ROC_{p_i} — коефіцієнт центральності вершини введений в [4]. Вартість переходу визначимо як суму вартостей перегляду кожної сторінки.

$$p_{ij} = p_i + p_j. \tag{2}$$

При такому підході інформаційна цінність інформаційної статті є величиною обернено пропорційною до вартості

$$v_i = \frac{1}{p_i}.$$

Запропонований підхід оцінки вартості враховує основні параметри поведінки користувача в гіпертексті. Основні припущення при цьому такі:

— якщо користувач відкрив сторінку, але фактично не переглядав її, вартість такого переходу збільшується;

— вартість залежить від типу сторінки — якщо сторінка є індексним показником, тобто містить значну кількість перехресних посилань, здійснюється зниження її вартості. З точки зору інформаційної цінності, цінність такої сторінки за рахунок введення індекс центральності збільшується;

— оскільки діапазон значень ROC_{p_i} залежить від величин перетвореної матриці відстаней, тобто є високим, замість безпосереднього значення індексу центральності використовується його логарифм;

— оскільки можлива гранична ситуація, у якій остання сторінка в ланцюгу перегляду є стоком, тобто не має гіперпосилань на інші сторінки, що призводить до $ROC_{p_i} = 0$, до під логарифмічного виразу додається 1.

Користувач гіпертексту в процесі навігації максимізує функцію корисності інформаційного перегляду

$$u = u\left(x_1,\ x_2, \ldots, x_N, I \right), \tag{3}$$

при вартісному обмеженні

$$\sum_i x_i p_i = I. \tag{4}$$

Визначимо лагранжіан L

$$L = u\left(x_1,\ x_2, \ldots, x_N,\ I \right) + \lambda \left(I - \sum_i p_i x_i \right), \tag{5}$$

де λ — множник Лагранжа, пов'язаний з (4). Тоді проводячи максимізацію, отримуємо, що мар-

шрут користувача у гіпертексті є розв'язком системи

$$\frac{\partial L}{\partial x_i} = \lambda p_i . \tag{6}$$

Цей розв'язок може бути записано у вигляді

$$x_i = x_i\left(p_1,\ p_2,\ldots, p_N, I\right), \tag{7}$$

і є функцією попиту на інформаційну сторінку i.

Із заданим рівнем корисності $\overline{u}$

$$\left.\frac{\partial I}{\partial p_i}\right|_{u=\overline{u}} = x_i . \tag{8}$$

Такі обчислення можна повторити для всього набору користувачів. Для користувача k рівняння (3)—(8) мають вигляд

$$u^{(k)} = u^{(k)}\left(x_1^{(k)},\ x_2^{(k)},\ldots,\ x_N^{(k)}, I^{(k)}\right);$$

$$\sum_i x_1^{(k)} p_i = I^{(k)};\quad \frac{\partial u^{(k)}}{\partial x_i^{(k)}} = \lambda^{(k)} p_i;\quad x_i^{(k)} = x_i^{(k)}\left(p_1,\ p_2,\ldots p_N,\ I^{(k)}\right);$$

$$\left.\frac{\partial I^{(k)}}{\partial p_i}\right|_{u^{(k)}=\overline{u}^{(k)}} = x_i^{(k)} . \tag{9}$$

Тепер можна визначити інформаційні витрати перегляду

$$I = \sum_k I^{(k)} ,$$

як сумарний витрачений час на перегляд гіпертекстової мережі користувачами, а попит на інформаційну одиницю i

$$x_i = \sum_k x_i^{(k)} .$$

За відповідних умов рівняння (9) можна додати по k. Тоді

$$\frac{\partial I}{\partial p_i} = x_i . \tag{10}$$

Рівняння (10) за формою запису збігається з рівнянням (8). Тому за відповідних умов рівняння (3)—(8) можуть бути інтерпретовані як рівняння, що описують одного користувача або групу користувачів в системі з максимальною корисністю.

Для запису ентропії введемо відповідні одиниці

$$y_i = \frac{x_i}{x}, \tag{11}$$

де

$$x = \sum_i x_i ; \tag{12}$$

$$S = -\sum_i y_i \ln y_i . \tag{13}$$

Відповідна фіксована величина в цьому випадку є I. Тоді визначимо

$$y_i = \frac{x_i p_i}{I}$$

як частину витрат, яку користувач витрачає при перегляді маршруту i. Тепер рівняння (13) може служити мірою ентропії. В цьому випадку система з максимальною корисністю може бути описана в термінах y_i таким чином:

$$u = u\left(\frac{y_1 I}{p_1},\ \frac{y_2 I}{p_2},\ldots,\ \frac{y_N I}{p_N}, I\right) \to \max . \tag{14}$$

$$\sum_i y_i = 1. \tag{15}$$

Функція Лагранжа має вигляд

$$L = u + \lambda\left(1 - \sum_i y_i\right). \tag{16}$$

Диференціюючи L по y_i отримаємо систему рівнянь:

$$\frac{\partial u}{\partial y_i} = \lambda, \quad i = 1,2,\ldots,N; \tag{17}$$

$$y_i = y_i\left(p_1,\ p_2,\ldots,p_N, I\right) \tag{18}$$

та

$$\left.\frac{\partial I}{\partial p_i}\right|_{u=\bar{u}} = \frac{y_i I}{p_i}. \tag{19}$$

Для аналізу цієї системи з використанням методу максимізації ентропії виділимо обмеження вигляду

$$f_k\left(y_1, y_2,\ldots, y_N\right) = g_k, \quad k = 1,\ldots,K, \tag{20}$$

де всі члени, що містять y_i входять до f_k, а всі інші — константні — до g_k.

Будемо розглядати лише сторінки для яких значення p_i задовольняє $p_i > M[P] + \sigma$, де $M[P]$ — математичне сподівання дискретної величини $P = \left(p_i \middle| i = 1,\ldots,N\right)$ за умови, її рівноймовірності. σ — середньоквадратичне відхилення. Таким чином в статті пропонується визначення складових маршрутів, що викликають найбільшу цікавість користувачів.

Нехай $V = \left(p_i \middle| p_i > M[P] + \sigma, i = 1,\ldots,N\right)$, тоді для гіпертекстової системи пропонується використання такого обмеження:

$$\frac{1}{N}\sum_{i=1}^{N} p_i = \frac{1}{|V|}\sum_{j=1}^{|V|} v_j. \tag{21}$$

Будемо максимізувати ентропію S, що визначена рівнянням (13) при обмеженнях (15) і (20). Це означає, що сформувавши лагранжіан у вигляді

$$L = S + \lambda\left(1 - \sum_i y_i\right) + \sum_k \mu_k \left(g_k - f\right)_k,$$

необхідно розв'язати систему рівнянь $\frac{\partial L}{\partial y_i} = 0$ разом з рівняннями (15) і (20).

Маємо

$$\ln y_i = -\lambda - \sum_k \mu_k \frac{\partial f_k}{\partial y_i} = 0,$$

і, вводячи одиницю в λ, отримуємо:

$$\ln y_i = -\lambda - \sum_k \mu_k \frac{\partial f_k}{\partial y_i}. \tag{22}$$

Нехай всі y_i, p_i та I були визначені правильно і обмеження (20) також були визначені коректно, тобто y_i у рівнянні (22) дають правильну модель, яка узгоджується з дійсністю. Тоді задача максимізації ентропії формально еквівалентна максимізації функції

$$u = S + \sum_k \mu_k \left(g_k - f_k\right) \tag{23}$$

з обмеженням (15). Тому, якщо система, що розглядається, є системою з максимальною корисністю, то можна вважати, що функція u, що визначається з (23) є функцією корисності. З обмеженням (21) маємо

$$u = S - \mu\left(\frac{1}{N}\sum_{i=1}^{N} p_i - \frac{1}{|V|}\sum_{j=1}^{|V|} v_j\right) \tag{24}$$

або

$$u = -\sum_i y_i \ln y_i - \mu\left(\frac{1}{N}\sum_{i=1}^{N} p_i - \frac{1}{|V|}\sum_{j=1}^{|V|} v_j\right). \tag{25}$$

Дослідження адекватності моделі

Дослідження адекватності розробленої моделі максимізації функції корисності користувача, що лежить в основі методу побудови оптимальних маршрутів у гіпертексті виконаємо експериментальним чином, тобто шляхом перевірки збігу моделі системи, що моделюється у відношенні цілі моделювання — мінімізації вартості перегляду маршруту визначеного в (1). Передбачається, що максимум функції корисності маршруту відповідає мінімуму вартості його перегляду та максимуму середнього часу, проведеного користувачем на кожній сторінці, що входить до нього.

Для проведення експерименту розроблена тестова гіпертекстова мережа, структура якої зображена на рис. 2. З метою отримання експериментальних даних сформовано групу користувачів, здійснено її розбиття на підгрупи, учасникам яких була надана інформація про матеріал, який їм необхідно знайти у цій мережі шляхом її послідовного перегляду. Тобто, учасники кожної підгрупи здійснювали незалежний пошук інформації зазначеної у завданні. Вибір початкової вершини перегляду здійснювався користувачами випадковим чином.

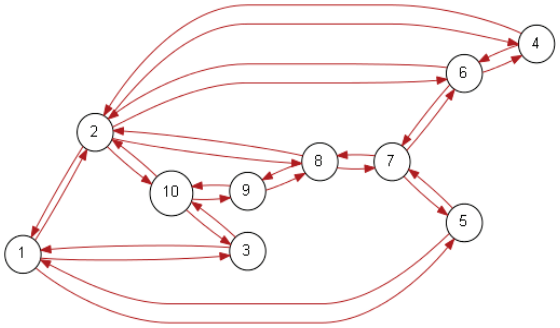

Рис. 2. Зображення сегменту гіпертекстової мережі, вибраної для проведення експерименту

pages =

	0	1	2	3	4	5	6	7	8	9
0	10	4	2	23	4	1	0	8	23	2
1	21	2	25	13	31	10	15	24	22	7
2	5	5	1	18	14	20	5	13	9	4
3	17	4	5	3	16	20	2	15	10	3
4	32	11	20	7	10	18	4	12	9	4
5	16	6	22	8	15	0	7	12	9	6
6	19	1	6	6	4	13	1	7	5	2
7	10	21	0	14	22	9	11	2	12	19
8	19	15	22	12	17	14	17	4	6	20
9	18	14	13	14	35	9	10	2	15	2

Рис. 3. Результати відвідування користувачами тестового сегменту гіпертекстової мережі

Результати відвідування користувачами сторінок мережі показані на рис. 3. Відповідно до даних результатів, можна зробити висновок, що найбільший потік користувачів виник між сторінками $p_5 \to p_{10}$, $p_5 \to p_1$, $p_2 \to p_3$.

Розглянемо приклад застосування моделі для випадку потоку користувачів $p_5 \to p_{10}$. Деталізація фактичних маршрутів користувачів між зазначеними сторінками вказана на рис. 4.

routes =

	0	1	2	3	4	5
0	1	2	0	0	0	0
1	1	6	7	2	8	9
2	7	8	2	0	0	0
3	1	6	4	2	8	9
4	7	6	1	2	0	0

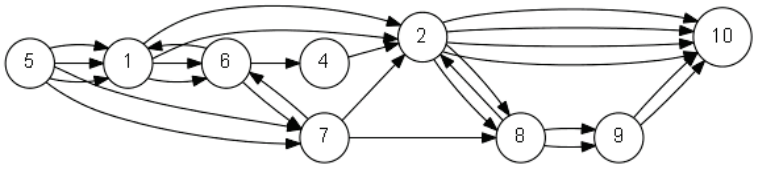

Рис. 4. Проміжні маршрути користувачів між сторінками 5 і 10

Відповідно користувачі витратили час у секундах, зазначений на рис. 5.

timespent=

	0	1	2	3	4	5
0	15	17	0	0	0	0
1	25	30	60	25	30	50
2	55	44	27	0	0	0
3	20	10	7	20	30	10
4	10	9	15	30	0	0

Рис. 5. Час, витрачений користувачами при навігації від вершини 5...10

result = $\begin{pmatrix} -4.261 \\ -2.207 \\ -0.308 \\ -25.984 \\ -16.457 \end{pmatrix}$

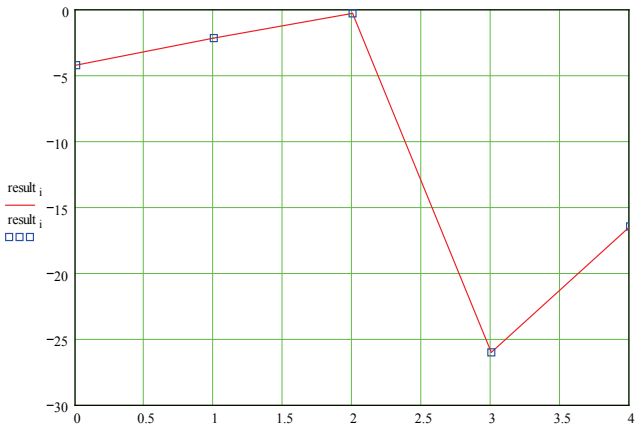

Рис. 6. Результат визначення функції корисності для маршрутів 0—4

Графічно результат визначення значень функції корисності для маршрутів показаний на рис. 6. Максимуму значення функція корисності набуває для маршруту № 2, що відповідає комбінації сторінок {7, 8, 2}. Як видно з результатів експерименту (26), маршрут № 2 характеризується найменшим значенням вартості (*costs*) — 3,026 та максимумом середнього часу перебування на сторінці (*avg_time*) — 42, що відповідає визначеним критеріям адекватності.

$$avg\_time = \begin{pmatrix} 16 \\ 36,667 \\ 42 \\ 16,167 \\ 16 \end{pmatrix}; \qquad \cos ts = \begin{pmatrix} 4,782 \\ 7,314 \\ 3,026 \\ 19,141 \\ 12,22 \end{pmatrix}. \tag{26}$$

Висновки

Запропонована модель поведінки користувачів у гіпертекстових мережах представляє собою засіб формалізації, що має у своїй основі ідентифікацію маршрутів із оптимальними значеннями вартості інформаційного перегляду, їх цінності та часу перебування користувача на сторінках, що входять до них.

Проведені дослідження підтверджують адекватність моделі та її відповідність цілям моделювання поведінки користувачів.

Запропонована модель є основою для розробки засобів оптимізації гіпертекстової структури на основі даних перегляду інформаційних сторінок користувачами, а також дозволяє автоматизувати цей процес.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Zatuchin D. Problem of website structure discovery and quality valuation / D. Zatuchin // Computer Science and Information Systems (FedCSIS), 2011 Federated Conference. — 2011. — P. 117—122.

2. Andaur E. Best Web Site Structure for Users Based on a Genetic Algorithm Approach [Електронний ресурс] / E. Andaur, S. Rios, P. E. Román and J. D. Velásquez // Procs. Of the First Workshop in Business Analytics and Optimization (BAO), Santiago, Chile. — 2010. — Режим доступу до журн. : http://wi.dii.uchile.cl/publications/conferences/Andaur2010.pdf.

3. Вилсон А. Д. Энтропийные методы моделирования сложных систем / А. Д. Вилсон. — М. : Наука, 1978. — 248 с.

4. Botafogo R. A. Identifying hierarchies and useful metrics / E. Rivlin, B. Shneiderman // ACM Transactions on Information Systems (TOIS). — 1992. — № 2. — P. 142—180.

Рекомендована кафедрою комп'ютерних систем управління

Стаття надійшла до редакції 18.10.11
Рекомендована до друку 10.11.11

Дубовой Володимир Михайлович — завідувач кафедри, ***Москвін Олексій Михайлович*** — аспірант.

Кафедра комп'ютерних систем управління, Вінницький національний технічний університет, Вінниця