УДК 517.977.5

В. М. Дубовой, д. т. н., проф.; О. М. Москвін

РОЗРОБКА НЕЧІТКОЇ СИСТЕМИ КЛАСИФІКАЦІЇ ГІПЕРТЕКСТОВИХ СТРУКТУР

У статті запропоновано нечітку систему оцінки оптимальності гіпертекстових інформаційних структур. Досліджено існуючі метрики для формалізації функціонально значущих їхніх параметрів на основі даних, отриманих в процесі дослідження впливу виду гіпертекстових структур на значення показників їхньої оцінки.

***Ключові слова**: гіпертекст, гіпертекстова метрика, індекс компактності, індекс стратифікації, нечітка класифікація, оптимальність структури.*

Вступ

Гіпертекстові інформаційні системи надзвичайно поширені, оскільки вони є основою Інтернет-ресурсів. Мережа гіпертекстових фрагментів характеризується складною, хаотичною та неоптимізованою структурою, що значно ускладнює пошук необхідної інформації. **Проблема** об'єктивного оцінювання характеристик гіпертексту з метою його оптимізації з розвитком Інтернет-технологій набуває все більшої **актуальності**.

У теорії гіпертексту для формалізації його функціонально значущих параметрів є спеціальна гіпертекстова метрика [1], яка містить два оцінних показники – індекс інформаційної компактності та індекс стратифікації. Як модель гіпертекстового інформаційного середовища обрано орієнтований граф, в якому вершинами є відповідні фрагменти, а ребрами – зв'язки між ними.

Індекс інформаційної компактності характеризує ступінь перетину гіпертекстової структури зв'язками [2]:

$$Cp = \frac{CD_{\max} - CD}{CD_{\max} - CD_{\min}}, \qquad (1)$$

де $CD_{\max}$ – максимально можлива кількість кроків, які необхідно пройти за посиланнями, що зв'язують усі вузли гіпертексту; $CD_{\min}$ – мінімально можлива кількість кроків, які зв'язують усі вузли гіпертексту; CD – показник шляхів у графі, для визначення якого необхідний попередній розрахунок перетвореної матриці відстаней.

Значення індексу інформаційної компактності варіює в межах [0; 1], що допускає порівняння систем гіпертекстових документів між собою. Абсолютно незв'язаний гіпертекст має індекс інформаційної компактності $Cp=0$, і навпаки, абсолютно зв'язний – $Cp=1$. Високий рівень компактності характеризує такі гіпертекстові структури, в яких на будь-який з інформаційних блоків можна з легкістю потрапити із будь-якого іншого блоку. Це зазвичай, забезпечується численними перехресними зв'язками. Потрібно зазначити, що надмірно висока компактність може призвести до повної дезорієнтації користувача гіпертекстової системи. У свою чергу, низька інформаційна компактність сприяє випаданню із поля зору багатьох фрагментів гіпертексту або призводить до втрати досяжності окремих фрагментів.

Індекс стратифікації детально розглянутий в [1] і введений для характеристики лінійності гіпертексту [3]:

$$St = \frac{AP}{LAP}, \qquad (2)$$

де AP – абсолютна стратифікація, а LAP – лінійна абсолютна стратифікація гіпертексту з

n вузлами ідентична до абсолютної стратифікації лінійного гіпертексту з *n* вузлами. Розраховується за формулою

$$LAP = \begin{cases} \dfrac{n^3}{4}, & \textit{якщо n парне} \\ \dfrac{n^3 - n}{4}, & \textit{якщо n непарне.} \end{cases} . \quad (3)$$

У випадку абсолютно стратифікованого гіпертексту, індекс стратифікації приймає значення *St=1* і ,навпаки, – *St=0*. Фактично, індекс стратифікації дозволяє оцінити рівень зв'язності елементів, які стоять на різних рівнях ієрархії.

Частка відсутніх шляхів, яка за змістом пов'язана з індексом інформаційної компактності:

$$K_m = \frac{Q_m}{n^2 - n}, \quad (4)$$

де Q_m – кількість відсутніх у графі шляхів [3]. Для розрахунку коефіцієнта відсутніх шляхів у графі необхідним є попереднє визначення матриці відстаней.

Максимальна кількість відсутніх шляхів дорівнює $n^2 - n$, мінімальна – 0. Значення частки відсутніх шляхів варіює в межах [0; 1] і допускає порівняння систем гіпертекстових документів між собою.

Цикломатичне число характеризує відмінність структури графа від деревоподібної структури і визначається за формулою:

$$Cn = m(G) - n(G) + p, \quad (5)$$

де m(G) – число ребер, n(G) – число вершин, p – число зв'язних компонентів графа [3].

Цикломатичне число показує найменшу кількість ребер, які необхідно видалити, щоб граф став деревом. Для сильнозв'язаного графа $p = 1$.

Аналіз критеріїв оцінки гіпертекстової структури показав неуніверсальність та неможливість їхнього окремого використання для отримання адекватної характеристики структури в зв'язку з їхнього функціональною обмеженістю.

Прикладом, який ілюструє недостатність кожного окремого показника для оцінювання якості гіпертексту є результати, зображені на рис. 1. Індекс стратифікації залишається однаковим як для деревоподібної структури без внутрішньо-ієрархічних зв'язків, так і з ними. З іншої сторони, значення індексу компактності як для лінійної замкненої структури так і для ієрархічної є майже однаковим.

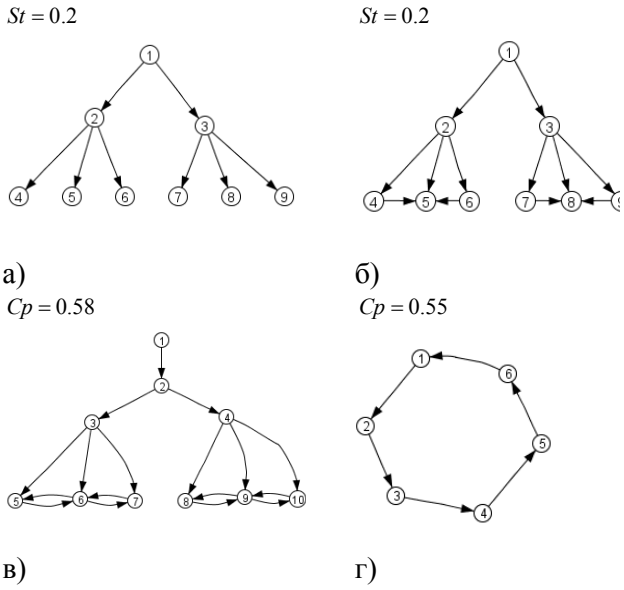


Рис. 1. Дослідження залежності виду структур від значення їхніх показників: а, б – індексу стратифікації; в, г – індексу інформаційної компактності

Завданням роботи є побудова системи нечіткої оцінки якості гіпертекстових структур як один з шляхів комплексного застосування існуючих показників.

Для оцінки впливу структури гіпертексту на значення критеріїв, були проведені дослідження з видозмінення гіпертекстових ієрархічних структур і обчислення значень розглянутих вище показників.

Базова ієрархічна структура, над якою проводилися операції модифікації зв'язків, є деревом і містить лише однонапрямлені зв'язки. Ця структура характеризується чіткою стратифікацією і має вигляд, представлений на рис. 2. Дослідження проводились ітераційним шляхом, під час якого на кожному кроці до графа за допомогою розробленого генератора ієрархічних структур додавалось по одному ребру та відбувалось обчислення значень індексу інформаційної компактності, індексу стратифікації, частки відсутніх шляхів та цикломатичного числа графа.

Проведено дослідження впливу різного типу зв'язків на значення критеріїв оцінки гіпертекстової структури:

- двосторонніх зв'язків між сусідніми рівнями;
- однонапрямлених зв'язків з нижчих рівнів до рівнів вищої ієрархії;
- однонапрямлених зв'язків з вищих рівнів до рівнів нижчої ієрархії;
- горизонтальних зв'язків між елементами окремих субієрархій;
- горизонтальних зв'язків між елементами різних субієрархій.

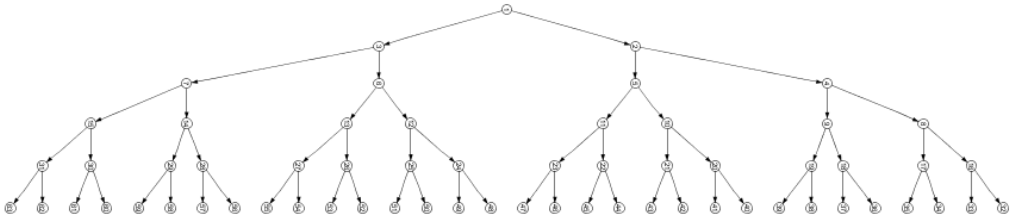


Рис. 2. Вид базової ієрархічної структури

Графічно принцип зміни графової структури, який використовується в цьому дослідженні, зображений на рис. 3. З нижніх рівнів додаються зворотні зв'язки до всіх вищепоставлених вершин у кожній субієрархії. Ці зв'язки позначені на рис. 3 штрих-пунктирними лініями.

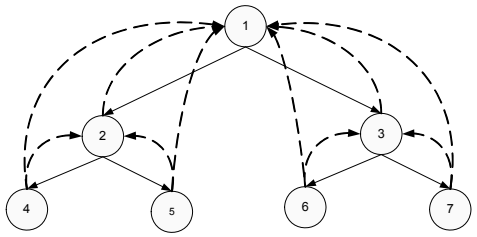
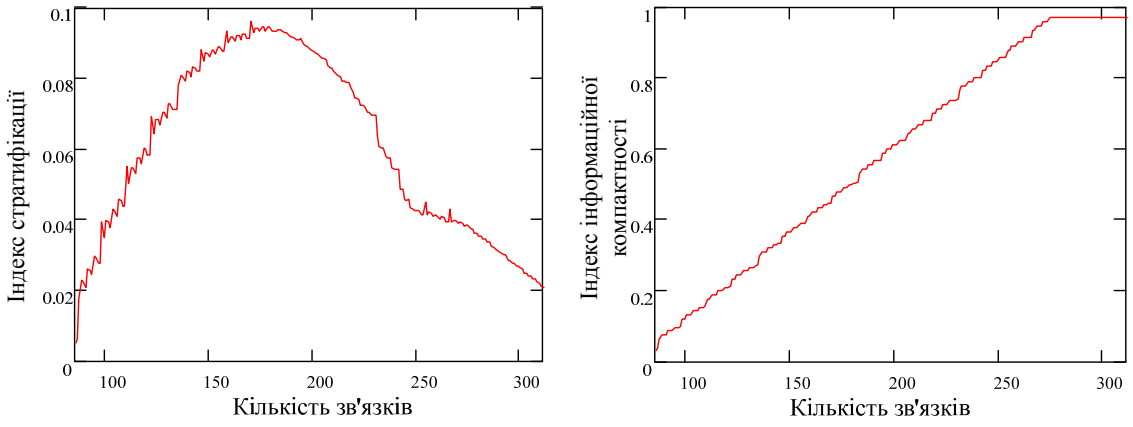


Рис. 3. Графічна інтерпретація принципу видозміни структури графа

Як видно з результатів досліджень (рис. 4), зміна значення індексу стратифікації має нелінійний характер. Початкове збільшення значення індексу стратифікації обумовлене появою зв'язків між кінцевими (що не мають вихідних зв'язків) і вищими за ієрархією вершинами. Повільний спад індексу стратифікації зумовлений ускладненням структури перехресними посиланнями.



а) б)

Рис. 4. Графічні залежності зміни критеріїв оцінки гіпертекстової структури від кількості однонапрямлених зв'язків типу "знизу-догори" у графі :
а) індексу стратифікації; б) індексу інформаційної компактності

Проведені дослідження дозволяють встановити евристичний зв'язок між змінами окремих характеристик гіпертексту і його якістю:

1. Зміна гіпертекстової структури шляхом додавання зворотних зв'язків тільки на 1 рівень вище надає можливість користувачеві вертатись назад у процесі навігації, але лише на один крок. Даний захід дає можливість зробити гіпертекст досяжним з будь-якої вершини. Виокремлене використання тільки цього підходу є невдалим, оскільки передбачає велику кількість переходів користувача між вузлами в процесі пошуку інформації. Більш прийнятним підходом до модифікації гіпертекстової структури є такий, в якому, крім зворотних зв'язків, додаються однонапрямлені зв'язки до рівнів вищої ієрархії. Швидкість досяжності фрагментів у такій структурі буде вищою. Оптимальною складністю такої структури будемо вважати таку, яка відповідає точці перегину характеристики залежності індексу стратифікації до кількості зв'язків, оскільки це є початком виродження структури.

2. Введення однонапрямлених зв'язків з вищих рівнів до рівнів нижчої ієрархії, як показали дослідження, є стратифікацію і майже не впливає на зміну індексу інформаційної компактності. Отже, виокремлене використання тільки цього підходу є неефективним, оскільки хоча і збільшується швидкість досяжності з верхніх рівнів ієрархії на нижні, верхні рівні фактично перевантажуються гіперпосиланнями, не надаючи користувачам можливості повторного повернення до них.

3. Горизонтальні зв'язки між елементами різних субієрархій незначно впливають на показник стратифікації, хоча істотно впливають на індекс інформаційної компактності. Дійсно, швидкість досягнення фрагментів значно збільшується за рахунок появи зв'язків з різними ієрархіями. Оптимальною складністю такої структури будемо вважати таку, яка відповідає точці перегину характеристики залежності індексу стратифікації до кількості зв'язків.

Узагальнимо отримані результати у вигляді нечіткої системи оцінювання якості гіпертекстової структури. Використання нечіткої логіки для цієї задачі є виправданим, оскільки отримані залежності виду гіпертекстової структури на відміну від значень її показників мають складний нелінійний характер. Пошук аналітичних залежностей, які їх описують, є проблематичним, а оціночні параметри в деяких випадках можуть давати суперечливі результати.

Дослідження і розробка нечіткої системи проводилась у пакеті прикладних програм Matlab 6, за допомогою інструментарію Fuzzy Logic Toolbox.

Згідно результатів досліджень була сформована нечітка база знань класифікації результатів оптимізації. Для створення системи нечіткого логічного виведення визначені 3 лінгвістичні змінні – індекс стратифікації, індекс інформаційної компактності та частка відсутніх шляхів.

Для системи нечіткого логічного виведення пропонується використання системи типу Мамдані, в якій значення вхідних і вихідної змінних задаються нечіткими термами [4].

База знань складається з 19 нечітких правил, які наведені у таблиці 1.

Візуалізація поверхні "входи-виходи" проведена за допомогою модуля Surface Viewer пакету прикладних програм Matlab.

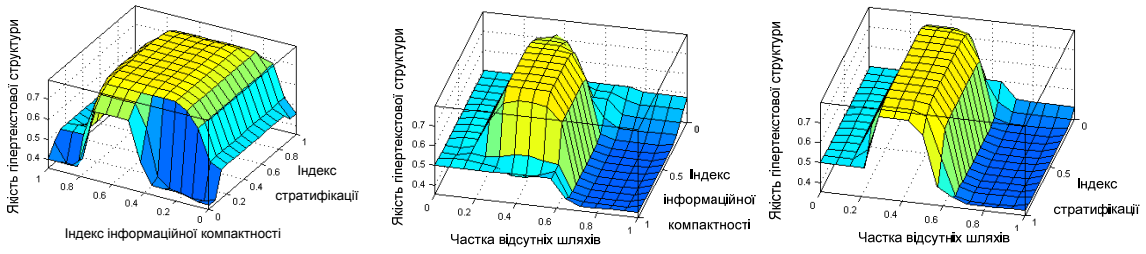
Таблиця 1

Правила нечіткої бази знань

Індекс стратифі-кації	Індекс інформаційної компактності	Частка відсутніх шляхів	Якість гіпертексто вої структури
високий	високий	високий	низька
високий	нижче середн.	середній	середня
середній	високий	низький	низька
середній	високий	низький	висока
середній	високий	середній	низька
середній	вище середн.	низький	середня
середній	вище середн.	середній	середня
середній	вище середн.	середній	висока
середній	нижче середн.	низький	низька
середній	нижче середн.	середній	середня
середній	середній	високий	низька
середній	середній	низький	низька
середній	середній	низький	висока
середній	середній	середній	висока
низький	високий	низький	низька
низький	високий	середній	низька
низький	нижче середн.	середній	низька
низький	низький	низький	низька
низький	низький	середній	низька

На рис. 5 зображені поверхні "входи-виходи" для вихідної змінної від комбінацій вхідних змінних – індекс стратифікації, індекс інформаційної компактності та частка відсутніх шляхів. Згідно з отриманими результатами, терму "високий" функції належності вихідної змінної відповідає певна опукла сфера.

Візуалізації, представлені на рис. 5, підтверджують, що сфера оптимальних розв'язків представляє собою певну їхню множину. Діапазони зміни значень параметрів оцінки для сфера оптимальних розв'язків згідно з результатами представленими на рис. 5 (а, б, в) є наступними: [0.2; 0.85] – для індексу інформаційної компактності, [0.1; 0.8] – для індексу стратифікації та [0.25; 0.6] – для частки відсутніх шляхів.



а) б) в)

Рис. 5. Поверхня "входи-вихід" у SurfaceViewer: а) входи – індекс інформаційної компактності, індекс стратифікації; б) входи – частка відсутніх шляхів, індекс стратифікації; в) входи – частка відсутніх шляхів, індекс інформаційної компактності

Висновок. Розроблена система нечіткої класифікації семантичної структури

Інтернет-ресурсів на основі даних 3-х показників (індексу інформаційної компактності, індексу стратифікації та частки відсутніх шляхів) дозволяє врахувати особливості кожного показника під час класифікації структури. Розроблені правила можуть буди використані для побудови автоматизованої системи оцінки семантичної структури сайтів за допомогою інструментарію FuzzyJ, який представляє собою набір бібліотек, які реалізують механізми нечіткої логіки для мови Java .

СПИСОК ЛІТЕРАТУРИ

1. The Semantic Web [Електронний ресурс] / Tim Berners-Lee, James Hendler, Ora Lassila // Scientific American Magazine. – May, 2001. – Режим доступу до журн.: http://www.sciam.com/article.cfm?id=00048144-10D2-1C70-84A9809EC588EF21.

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

3. Harary F. Structural models. An Introduction to the Theory of Directed Graphs / Harary F., Norman R., Cartwright D. – Wiley: New York, 1965. – 415 p.

4. Ротштейн О. П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети / Ротштейн О. П. – Вінниця: «УНІВЕРСУМ-Вінниця», 1999. – 320 с.

Дубовой Володимир Михайлович – д. т. н., професор, завідувач кафедри комп'ютерних систем управління. тел.: (0432) 598-157, E-Mail: dub@faksu.vstu.vinnica.ua.

Москвін Олексій Михайлович – аспірант кафедри комп'ютерних систем управління.
Вінницький національний технічний університет.