УДК 681.3; 519.688

П.Ф. Павлов
Харківський Національний Університет радіоелектроніки,
науково-учбова лабораторія «Придбання знань»

МАТЕМАТИЧНІ МОДЕЛІ ГІПЕРТЕКСТОВИХ СТРУКТУР

This paper considers mathematical models of hypertextical structures. Hypertextical structure of encyclopedia are showed. The coefficient deviation from linear sequence are propose for hypertext.

У статті розглядаються математичні моделі гіпертекстових структур. Показана гіпертекстова структура енциклопедії. Запропоновано коефіцієнт відхилення гіпертексту від лінійної послідовності.

Вступ

Майже кожна сучасна освічена людина знає слово "*гіпертекст*", але ще кілька років тому з цим терміном були знайомі лише фахівці з інформаційних технологій опрацювання природномовних текстів.

"Термін *гіпертекст* був введений у звертання Тедом Нельсоном (Ted Nelson) у 1965 році для опису документів (наприклад, тих які представлені у комп'ютері), що виражають нелінійну структуру ідей, на противагу лінійній структурі традиційних книг, фільмів і мови. Пізніший термін "*гіпермедіа*" близький до нього за змістом, але він підкреслює наявність у *гіпертексті* нетекстових компонентів, таких як анімація, записаний звук і відео" [1].

Та інформаційна структура, яку Нельсон запропонував називати гіпертекстом не є чимось новим. Так засіб подання інформації, що здавна використовується в енциклопедіях, енциклопедичних словниках, тлумачних словниках та ін., тобто у тих видах видань, де текст розбито на окремі статті (обсягом від кількох слів у тлумачних словниках до кількох сторінок в енциклопедіях), а самі статті мають (за звичай) посилання на інші статті цього ж видання, також можна вважати за гіпертекст. Наявність посилань (зв'язків між статтями) є суттєвою властивістю цього виду видань, яка відрізняє його від таких видів видань, як збірники статей, часописи та газети, які теж складаються з окремих статей, але таких посилань не містять. Збірник наукових статей, наприклад [2], складається з окремих статей. Ці статті мають посилання на інші джерела (монографії, статті з інших збірників та часописів, та т.п.). Слід зазначити, що посилання на інші джерела є характерною рисою наукових видів видань, хоча і інші види видань теж використовують посилання на інші джерела.

Отже для наукових статей характерним є наявність посилань на інші джерела (наявність зовнішніх посилань), які вже існують, тобто виникли раніше за ту статтю, що має ці посилання. Безумовно, бувають такі випадки, коли стаття у збірнику посилається на іншу

статтю у цьому ж збірнику. Але це виняток, який виникає в ситуації, коли автор (колектив авторів) з якихось міркувань подає цілісну роботу як сукупність кількох статей.

Енциклопедія як приклад гіпертексту

Більш докладно розглянемо організацію енциклопедії. Зазначимо насамперед, що слово "енциклопедія" перекладається з грецької як "навчання по всьому колу знань". З цього визначення видно, по-перше, що енциклопедія призначена для процесу навчання, тобто формування системи знань людини, по-друге, що енциклопедія повинна містити всі знання. Зрозуміло, що з плином часу попереднє значення набуло нового змісту. Зараз під енциклопедією розуміється видання наукового або науково-популярного характеру, що містить систематизоване та грунтовне зведення знань. За тематикою енциклопедії розподіляють на загальні, спеціалізовані (галузеві) та регіональні. Із загальних

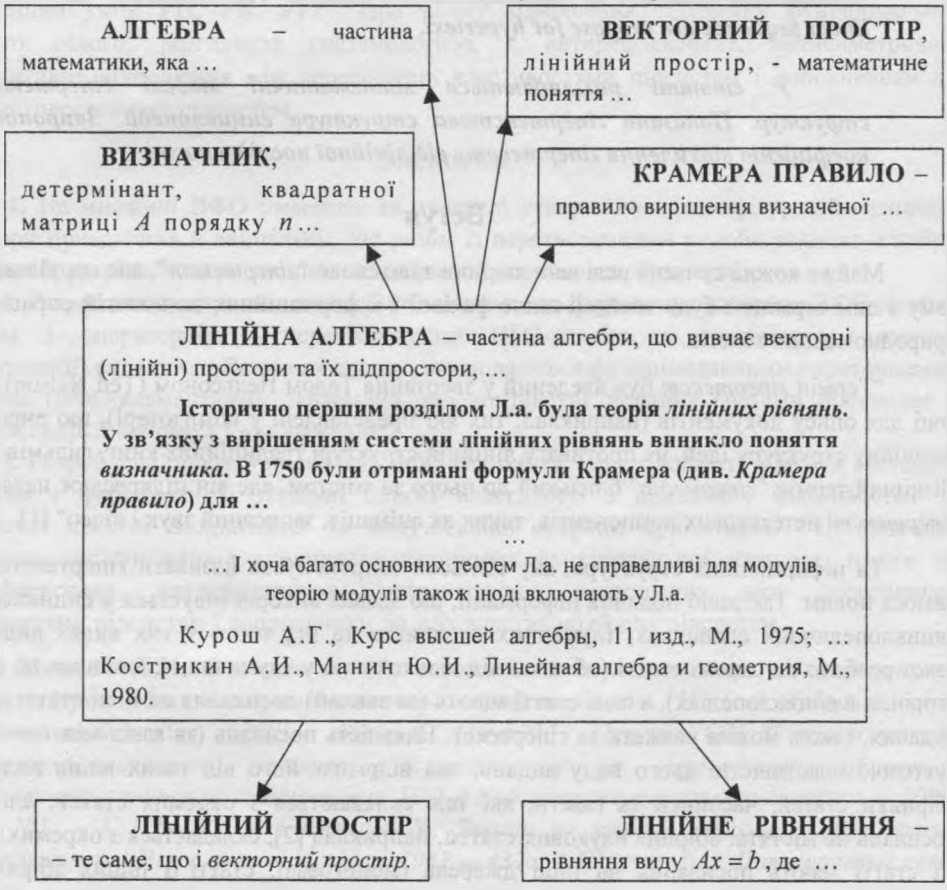

Рис. 1. *Схема зв'язків енциклопедичної статті "Лінійна алгебра" з іншими статтями енциклопедії*

енциклопедій найвідомішою є Британіка (Encyclopedia Britannica). Крім терміну

"енциклопедія" зараз вжитку є похідні від нього терміни "енциклопедичні знання" (глибокі та вичерпні знання) та "енциклопедист" (людина, що володіє енциклопедичними знаннями).

Статті в енциклопедіях складаються з назви статті та тексту статті. Крім того, в деяких енциклопедіях прийнято вказувати автора (авторів) статті та зовнішні джерела (наприклад [3]). Найцікавіше міститься в тексті статті, а саме посилання на інші статті енциклопедії. За звичай їх (посилання) подають як слова, що виділені або курсивом, або іншим способом. Інколи використовують і пряме звертання до іншої статті, даючи вказівки типу "дивись" + "назва статті".

Наприклад, на рисунку 1 наведено схематичне зображення статті "Лінійна алгебра" з математичного енциклопедичного словника [4], яка має посилання на інші статті цієї ж енциклопедії. Зокрема на статті "Лінійне рівняння", "Визначник" та "Крамера правило" посилання позначається виділенням назв цих статей курсивом. Крім того, на статтю "Крамера правило" дається пряме звертання за допомогою слова *дивись*. Назви інших статей, які пов'язані зі статтею "Лінійна алгебра" і зустрічаються у тексті самої статті (а це статті "Алгебра", "Векторний простір", "Лінійний простір", "Модуль" та "Підпростір") ні як не позначені. Ці зв'язки користувач повинен виводити сам, шукаючи відповідні терміни у енциклопедії.

Структуру наведеної на рис. 1 енциклопедичної статті можна записати як четвірку:

$E = < n, T, A, S >$,

де n – назва статті, T – текст статті, A – автор(и) статті, S – зовнішні джерела. Елемент n цієї четвірки можна розглядати трьома способами:

1) як якийсь елемент, що належить до множини всіх назв статей N;

2) як упорядкований набір, що складається з лексем, які належать до словника L, тобто ланцюжок з елементів $l_1 l_2 \ldots l_k$, де $l_i \in L$;

3) як ланцюжок символів $\omega_1 \omega_2 \ldots \omega_m$, що належать алфавітові Ω.

Елемент T четвірки E можна розглядати двома способами:

- як упорядкований набір лексем (подібно до способу 2 для елемента n);

- як вислів в алфавіті Ω (подібно до способу 3 для елемента n).

Елементи A та S четвірки E також можуть бути розглянуті як ланцюжок символів $\omega_1 \omega_2 \ldots \omega_m$, що належать алфавітові Ω. Цей спосіб є найнижчім, так би мовити "фізичним" способом подання елементів четвірки і спирається на саму природу інформаційних технологій. Будемо його називати символьним методом. Символьний метод повинен включати, як мінімум, визначення алфавіту та правил складання висловів у цьому алфавіті. У нашому випадку алфавіт Ω - це набір символів, що належать різним природномовним алфавітам, а також символи, що належать спеціалізованим алфавітам (математичні символи, символи для опису хімічних формул і тому подібне). Правила для отримання "правильних" висловів у нашому випадку (особливо для елемента T) є доволі складними та громіздкими, не тільки для застосування в практиці, але і для теоретичного аналізу.

Лексичний метод, який було застосовано для елементів n та T (подання елементів як упорядкованого набору, що складається з лексем, які належать до словника L) можна застосувати, також, до елементів A та S четвірки E.

Введемо множину $Q = \{q_i\}$. Кожен елемент q цієї множини позначає одного автора. У даному випадку нема однозначної відповідності між поняттям автор та людини. Одній і тій же людині може відповідати кілька елементів у множині Q. Це відбувається коли людина публікується під різними своїми іменами (прізвище людини не є його сталою характеристикою, особливо у жінки), або людина з якихось причин використовує псевдоніми (наприклад, авторка популярних підручників з теорії ймовірностей доктор технічних наук Олена Сергіївна Вентцель друкувала свої художні твори під псевдонімом "Грекова И"). Бувають випадки, коли за одним ім'ям приховується цілий авторський колектив (наприклад відомий в математиці псевдонім Nicolas Bourbaki слугує цілій школі французьких математиків вже кілька десятків років). Але використання псевдонімів для наукових видань є не характерним явищем. За допомогою множини Q ми можемо визначити елемент A четвірки E як кортеж, тобто $A =< q_1, q_2, \ldots, q_k >$, де $q_i \in Q$. Випадок, коли арність кортежу дорівнює нулю ($k = 0$), відповідає ситуації коли авторство статті в енциклопедії не зазначено.

Введемо множину $F = \{f_i\}$. Кожен елемент f цієї множини позначає одне джерело, на яке може посилатись стаття енциклопедії. За допомогою множини F ми можемо визначити елемент S четвірки E як кортеж, тобто $S =< f_1, f_2, \ldots, f_k >$, де $f_i \in F$. Випадок, коли арність кортежу дорівнює нулю ($k = 0$), відповідає ситуації коли стаття в енциклопедії не має посилань на зовнішні джерела.

Розглянемо вище введену множину всіх назв статей $N = \{n_i\}$. Назва енциклопедичної статті n, що належить цій множині $n \in N$, може бути представлена як ланцюжок лексем визначених на множені L, тобто $n_j = l_{j1} l_{j2} \ldots l_{jk}$ де $l_{ji} \in L$. Таким чином множину всіх назв статей N можна задати як відношення на множині $L^m = \underbrace{L \times L \times \ldots \times L}_{m \text{ раз}}$, де m максимальна довжина ланцюжка для будь якого елемента n_j, що належать N. Через $L_N = \{l'_N\}$ позначимо множину, що складається з ланцюжків лексем довжиною від одного до m. Кожному елементу множини L_N відповідає один і тільки один елемент множини N, і навпаки, кожному елементу множини N відповідає один і тільки один елемент множини L_N, тобто між множинами L_N та N існує взаємно однозначна відповідність.

На множині L_N задамо порядок елементів. Спочатку елементи впорядкуємо за довжиною ланцюжка за зменшенням, а серед елементів однакової довжини використаємо лексикографічний порядок. Для завдання порядку на лексемах, кожна з яких є ланцюжком символів з алфавіті Ω, необхідно, спочатку, визначити порядок на алфавіті Ω, тобто для кожної пари елементів $\omega_i, \omega_j \in \Omega$, де $i \neq j$, визначити яке з порівнянь $\omega_i < \omega_j$ чи $\omega_j < \omega_i$ є вірним. Впорядкувавши алфавіт Ω, ми тим самим, отримаємо строгий порядок на множині

L_N. Тепер за допомогою алгоритму перебору ми можемо визначити, які ланцюжки лексем, що належать множині L_N зустрічаються в елементі T, який теж є ланцюжком, що складається з лексем. Якщо ланцюжок лексем $l_N^a \in L_N$, що відповідає назві енциклопедичної статті $a \in N$, зустрічається в елементі T^b, що належить четвірці енциклопедичної статті b: $E^b = < b, T^b, A^b, S^b >$, то це означає, що енциклопедична стаття b має посилання на енциклопедичну статтю a. Занотуємо цей факт як пару $< b,a >$, де $b,a \in N$, та будемо називати цю пару посиланням. Позначимо через R_b множину всіх посилань для статті b, а через $R = \bigcup_{b \in N} R_b$ – множину всіх посилань всіх енциклопедичних статей. Зрозуміло, що $R \subseteq N^2$.

Позначимо через V сукупність всіх статей енциклопедії $V = \{E_i\}$. В цьому разі енциклопедія може бути визначена як упорядкована пара з двох елементів: множини всіх енциклопедичних статей V та множини всіх посилань всіх енциклопедичних статей R, тобто $\Theta = < V, R >$. Таким чином, енциклопедія Θ, як інформаційна структура, може бути представлена як граф, де V - це множина вершин (вузлів), а R - це множина ребер (дуг). Слід зазначити, що це визначення енциклопедії з точністю до одного елемента (броузера) співпадає з визначенням тривіальної гіпертекстової системи введеному у [5].

Броузер – це механізм навігації на гіпертексті, який дозволяє переходити від одного вузла гіпертексту до іншого використовуючи існуючі між вузлами дуги. Зрозуміло, що енциклопедія немає такого механізму, бо не належить до автоматизованих систем, але тезу що енциклопедія має гіпертекстову структуру будемо вважати доведеною.

Гіпертекст та HTML

В радянській літературі одним з перших використав термін гіпертекст Суботін [6], який спирався на визначення гіпертексту запропоноване Колкіним [7] – "гіпертекст – це форма організації текстового матеріалу, за якої його одиниці представлені не в лінійній послідовності, а як система явно вказаних можливих переходів, зв'язків між ними. Слідуючи за цими зв'язками, можна читати матеріал у будь-якому порядку, утворюючи лінійні тексти. Якщо мова йде про достатньо широкий матеріал з великою кількістю зв'язків, то виникає доволі складний гіпертекстовий простір (мережа). Формування та проглядання такої мережі текстових одиниць можливе тільки за допомогою комп'ютера".

В сучасних роботах визначення гіпертексту дуже часто прив'язують до однієї з сфер його застосування. Для прикладу наведемо визначення гіпертексту з [8]: "Гіпертекст – це легка у використанні, але надзвичайно потужна система пов'язаних слів і фраз, яка дозволяє здійснювати навігацію між сторінками. Ці слова є перехресними посиланнями на інші слова на інших сторінках, і, зазвичай, виділяються на сторінці Web яскравішим кольором."

Безумовно, Інтернет, особливо з появоюWeb, став провідною технологією сучасної IT-індустрії. Одним з базових компонентів цієї технології є мова HTML (Hypertext Markup Language). З назви цієї мови видно, що вона призначена для створення гіпертексту. Це досягається шляхом використання управляючих маркерів – тегів (tag). Зокрема, тег "a"

використовується для вказівки посилання. Посилатись можна на маркер у самому тексті (в поточному), на інший текст (який знаходиться в іншому файлі, шлях до якого вказується), на маркер в іншому тексті.

В якості прикладу наведемо один абзац для енциклопедичної статті, що зображена на рис. 1:

```
<li> Історично першим розділом Л.а. була теорія <I><a
href=mencsl_313.htm>лінійних рівнянь</a></I>. У зв'язку з
вирішенням системи лінійних рівнянь виникло поняття <I><a
href=mencsl_168.htm>визначника</a></I>. В 1750 були отримані
формули Крамера (див. <I><a href=mencsl_284.htm>Крамера
правило</a></I>) для вирішення системи лінійних рівнянь, у якій
число рівнянь дорівнює числу невідомих і визначник із коефіцієнтів
при невідомих відмінний від нуля. </li>
```

В цьому прикладі зустрічається три посилання: на статтю "Лінійне рівняння", що зберігається у файлі `mencsl_313.htm`; на статтю "Визначник" (файл `mencsl_168.htm`) та на статтю "Крамера правило" (файл `mencsl_284.htm`).

В мережі Інтернет можна знайти багато документів (підручників, учбових посібників і т.п.), які складаються з кількох файлів у форматі HTML і мають структуру, що зображена на рис. 2.

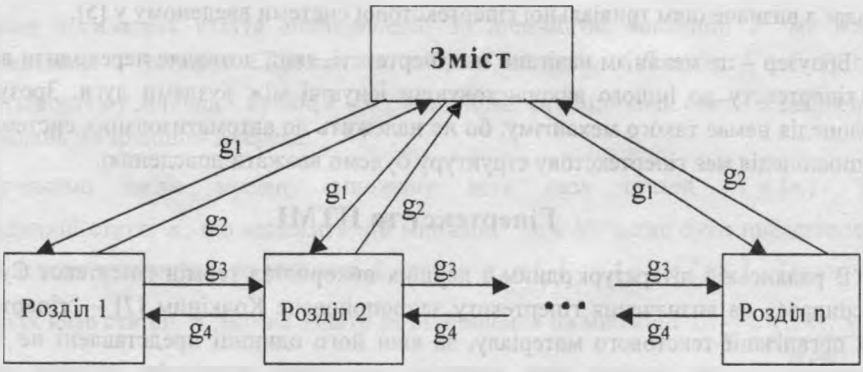

Рис. 2.

У цій структурі дуги мають таку семантику: g_1 - перехід від змісту до розділу книги; g_2 - перехід до змісту; g_3 - перехід до наступного розділу; g_4 - перехід до попереднього розділу. Ці чотири типи дуг утворюють множину $G=\{g_1,g_2,g_3,g_4\}$ - множину підтримки лінійної структури тексту. Завдяки цій множині можна ввести коефіцієнт h - відхилення гіпертексту від лінійної структури, як $h=\frac{Q-Q_G}{Q}$, де Q - загальна кількість дуг у гіпертексті, Q_G - кількість дуг у гіпертексті, що належить множині G. Зрозуміло, що гіпертекст зі значенням коефіцієнта $h=0$ важко вважати текстом в якому "одиниці представлені не в лінійній послідовності". Слід зазначити, що для більшості енциклопедій (тих, що не мають змісту) цей коефіцієнт буде дорівнювати одиниці.

Розглянемо довільний гіпертекст H, що складається з вершин $h_i \in H$. Якщо дві вершини h_α та h_β належать гіпертексту H і вершина h_α має тільки одну дугу типу $g_3^\alpha = < h_\alpha, h_\beta >$ та вершина h_β має тільки одну дугу типу $g_4^\beta = < h_\beta, h_\alpha >$, тоді можна ввести операцію повної лінійної згортки вершин h_α та h_β:

$$Z : < h_\alpha, h_\beta > \mapsto h_\gamma ,$$

де вершина h_γ має такі властивості – текст її є конкатенацією текстів вершин h_α та h_β, тобто $T_\gamma = T_\alpha T_\beta$, а множина її дуг складається з дуг, що належали вершинам h_α та h_β за винятком дуг g_3^α та g_4^β, тобто $R_\gamma = R_\alpha \cup R_\beta - \{g_3^\alpha, g_4^\beta\}$. В разі виконання операції повної лінійної згортки для вершин h_α та h_β, вони вилучаються з гіпертексту H, а вершина h_γ додається до H. Крім того, для всіх інших вершин гіпертексту повинна бути виконана заміна дуг, що посилались на вершини h_α та h_β, посиланнями на вершину h_γ. Очевидно, що для будь-якої вершини, яка належить гіпертексту може існувати тільки одна вершина, з якою можна утворити пару для виконання операції повної лінійної згортки.

Висновки

Завдяки проведеному аналізу структури енциклопедії показано, що гіпертекст не є якимось принципово новим явищем створеним сучасними інформаційними технологіями. Гіпертекст є скоріше новим підходом до використання інформаційних структур, які подібні до структури знань існуючої у людини. Цей підхід спирається на сучасні програмно-апаратні ресурси і технології. Завдяки подібності гіпертекстових структур до структури системи знань людини використання гіпертексту підвищує ефективність інтелектуальної діяльності людини.

На даний момент ще не склалася теорія гіпертексту і багато задач в цьому напрямку вимагають свого вирішення.

1. Microsoft Press. Толковый словарь по вычислительной технике/Пер. с англ. – М.: Издательский отдел "Русская редакция" ТОО "Channel Trading Ltd", 1995. – 496 с. 2. Інформаційні системи та мережі, Вісник ДУ «Львівська політехніка», №315. Львів, 1997. 3. Математическая энциклопедия, т. 1-5, М.: Сов. энциклопедия, 1977-85 4. Математический энциклопедический словарь, М.: Сов. энциклопедия, 1988. – 847 с. 5. Павлов П.Ф. Применение нечетких пространств толерантности для моделирования знаний на основе гипертекстовых структур // Проблемы бионики: Всеукр. межвед. науч.-техн. сб 2001. Вып. 55. С. 43-45. 6. Субботин М.М. Новая информационная технология: создание и обработка гипертекстов. // НТИ. - Сер. 2. - 1988. N 5. - с. 2-7. Conklin J. Hypertext: An introduction and survey // Computer. – 1987. – N 9. – P. 17-41. 8. Галайко В.М. Розроблення інтелектуальних Web-систем // Інформаційні системи та мережі, Вісник ДУ «Львівська політехніка», №330. Львів, 1998. – С. 52-62