Київська міська адміністрація
Інститут прикладної інформатики

На правах рукопису

ЖОГОВ Георгій Володимирович

АДАПТИВНИЙ ІНФОРМАЦІЙНИЙ ПОШУК. ПІДХІД І РЕАЛІЗАЦІЯ

05.13.17 - теоретичні основи інформатики

Автореферат дисертації на здобуття наукового ступеня
кандидата фізико-математичних наук

Науковий керівник - старший науковий співробітник
кандидат фізико-математичних наук
ДРІЯНСЬКИЙ В.М.

Київ - 1994

АВ 31.003

Дисертація є рукопис

Робота виконана в Інституті прикладної інформатики

Науковий керівник кандидат фізико-математичних наук, старший науковий співробітник Дрянський Володимир Михайлович

Офіційні опоненти:

1. Доктор фізико-математичних наук, член-кореспондент АН України, Ющенко Катерина Логвінівна

2. Кандидат фізико-математичних наук, старший науковий співробітник, Полуміенко Сергій Костянтинович

Провідна організація Інститут програмних систем НАН України (м. Київ)

Захист відбудеться "25" жовтня 1994 р. об 10 год. на засіданні Спеціалізованої вченої ради Д 166.01.01 в інституті прикладної інформатики (ІПрІн) за адресою: 252004, м. Київ, вул. Червоноармійська, 23-б.

З дисертацією можна ознайомитися у бібліотеці інституту прикладної інформатики.

Автореферат розісланий "23" вересня 1994 р.

Вчений секретар
Спеціалізованої вченої ради Мелент'єв Г.Б.

АВ-31.003

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. В інформаційних системах, орієнтованих на ефективну обробку великих об'ємів текстографічної інформації: бібліотечних, науково-технічних, юридичних, офісних и т.п., особливе значення набувають можливості адаптації пошукових механізмів до інформаційних інтересів конкретних абонентів, тобто наявність засобів, які забезпечують, по-перше, початкове інформаційне позиціонування користувача і створення його особистого інформаційного середовища і, по-друге, підтримку цього середовища в актуальному стані.

У відомих пошукових схемах для рішення цієї проблеми в значній мірі використовуються інтелектуальні зусилля абонентів пошукової системи. Ось чому створення нових пошукових моделей з елементами штучного інтелекту, які значно зменшують інтелектуальне навантаження абонента, є актуальним.

Ціль роботи - розробка схеми адаптивного інформаційного пошуку і визначення в ній процесів, які дозволяють використання адаптаційних механізмів; моделювання цих процесів, розробка алгоритмів функціонування побудованих моделей і створення прототипу відповідної інформаційної системи.

Наукова новина. Запропонована концепція адаптивного інформаційного пошуку і схема інформаційної технології, яка базується на цій концепції. В рамках цієї схеми побудовані моделі локалізації інформаційної потреби абонентів і вивчені її властивості, запропоновані семантичні розширення цих моделей, розроблені алгоритми їх функціонування. Створено прототип системи адаптивного інформаційного пошуку і здійснено цикл експериментів, які підтвердили правильність запропонованої концепції.

Загальна методика досліджень. Для моделювання процесів локалізації інформаційної потреби виділяються два рівня опису: міжсеансовий та внутрішньосеансовий. Моделі міжсеансового рівня побудовані з використанням алгебраїчного підходу до теорії розпізнавання образів (по Ю.І.Журавльову). Для внутрішньосеансового моделювання використовується байєсовський підхід у статистичній теорії прийняття рішень, а для побудови семантичних розширень - теорія відношень.

Практична цінність. Реализація концепції адаптивного пошуку

приводить до появи принципово нового класу інтелектуальних інформаційних систем з "прозорим" інтерфейсом користувача. Використання подібних систем дозволяє по-новому подивитись на цілий ряд галузей використання інформаційних технологій: перш за все це бібліографічні інформаційні системи, електронні каталоги і інформаційні мережи, оскільки в межах запропонованого підходу може бути здійснено автоматичне виявлення інформаційних потреб користувачів і подальше автоматичне вибіркове розповсюдження інформації в їх персональні бази даних (БД).

Реалізація результатів досліджень. Реалізовано і перевірено на експериментальній БД прототип системи адаптивного інформаційного пошуку. Система була запроваджена в Експо-центрі "Наука" АН України. Елементи цієї системи були використані при розробці експертної системи скринінгу хворих вторинним імунодефіцитом і системи обробки кореспонденції ДКНТ України. За цикл робіт "Настраивающиеся документальные системы: математические модели, алгоритмы, программное обеспечение" автор спільно з Л.Г.Катериничем та О.Я.Колтуном був удостоєний у 1986 р. медалі АН України з премією для молодих вчених.

Апробація роботи. Матеріали дисертації доповідались на семінарах Наукової ради АН України з проблеми "Кібернетика"; "Проблемы проектирования автоматизированных банков данных"; "Лингвистические проблемы проектирования информационных систем"; "Распознавание и оптимальное управление развитием систем" (Славское, 1989 р.); на Республіканській науково-технічній конференції "Пути дальнейшего совершенствования системы научно-технической информации и пропаганды в Украинской ССР в XII пятилетке" (Хмельницкий, 1986 р.); на Всесоюзному науково-технічному семінарі "Мобильное программное обеспечение" (Калінін, 1988 р.); на 5 Всесоюзній конференції по БД та знань (Львів, 1991 р.).

Публікації. За результатами виконаних досліджень опубліковано 11 робіт, одна колективна монографія (глави 6, 10).

Обсяг і структура роботи. Дисертація складається з вступу, трьох глав, заключення, списку літератури (63 найменування) і додатків. Дисертація викладена на 130 сторінках, містить 4 малюнка та 2 таблиці. Додатки на 31 сторінках.

ЗМІСТ РОБОТИ

Дисертаційна робота присвячена питанням розробки і реалізації схеми адаптивного інформаційного пошуку:

- моделюванню процесів локалізації і ведення інформаційної потреби користувача, а також власне інформаційного пошуку і оцінки знайденої інформації;

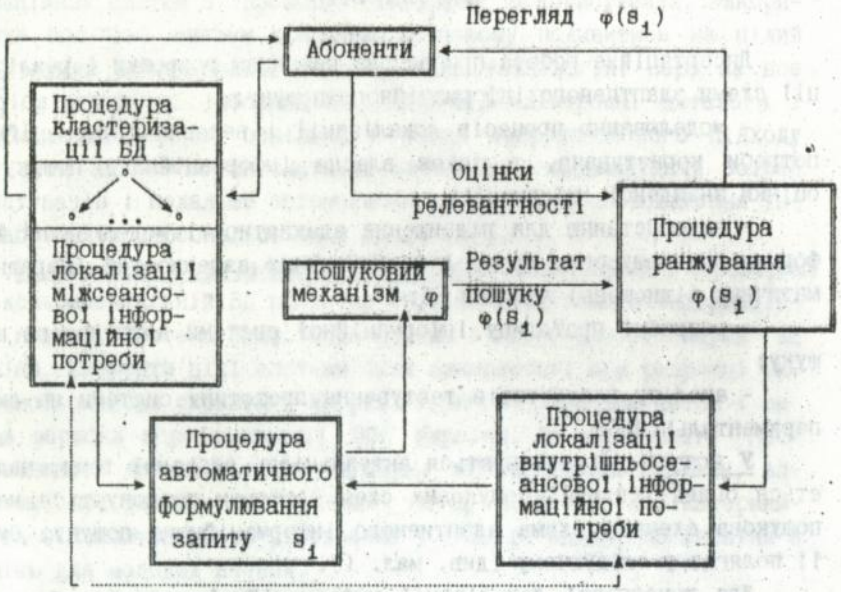
- використанню для підвищення адекватності моделювання інформаційно-пошукових процесів семантичних залежностей (парадигматичних відношень) лексики БД;

- розробці прототипу інформаційної системи адаптивного пошуку;

- аналізу результатів тестування прототипу системи на експериментальній БД.

У вступі обгрунтовується актуальність вибраної теми, надається огляд існуючих пошукових схем. Автором пропонується нова пошукова схема - схема адаптивного інформаційного пошуку. Суть її полягає у слідуючому (див. мал. 1).

Для початкової локалізації информаційної потреби абоненту надається кластерне дерево, яке відображає тематичне розподілення БД. В процесі його перегляду абонент визначає деяку підмножину кластерів, які його зацікавили. На базі цієї інформації автоматично формулюється запит s_0, по якому організовується ітераційний пошук. Тобто, s_0 подається на вхід пошукового механізму φ. Результати пошуку $\varphi(s_1)$ (в початковий момент i=0) проглядаються абонентом. Результати оцінок $\varphi(s_i)$ подаються на вхід процедури внутрішньосеансової локалізації інформаційної потреби, результати роботи якої визначають подальшу роботу процедури автоматичного формулювання запиту s_{i+1}. Новий запит знову подається на вхід пошукового механізму. Ітераційний цикл зупиняється або системою, або користувачем. Результати спілкування з системою запам'ятовуються для подальшого використання у процедурі міжсеансової локалізації інформаційної потреби, яка дозволяє, по-перше, відслідкувати динаміку змінювання інформаційної потреби абонента від сеанса до сеансу і, по-друге, впливати на роботу процедури внутрішньосеансової локалізації.



Малюнок 1. Схема адаптивного інформаційного пошуку

У першій главі описуються розроблені автором математичні моделі локалізації інформаційної потреби користувача в схемі адаптивного інформаційного пошуку. Для вирішення задачі локалізації інформаційної потреби на базі схеми адаптивного пошуку пропонується виділити два види інформаційної потреби,- міжсеансову і внутрішньосеансову. Локалізація міжсеансової інформаційної потреби спрямована на визначення динаміки змінювання інформаційних інтересів користувачей і використовується для вирішення задачі класифікації абонентів. Локалізація внутрішньосеансової інформаційної потреби необхідна для ефективної навігації абонента по БД в межах одного сеансу спілкування його з системою.

Визначення 1. Міжсеансовою інформаційною потребою будемо називати підмножину кластерів класифікаційної структури БД, які були визнані абонентом пертинентними в процесі її перегляду.

Очевидно, що задачу локалізації міжсеансової інформаційної потреби можна розглядати як класичну задачу розпізнавання, а са-

ме, по опису множини документів $K = K_1 \oplus K_2$, про яку відомо, що документи з K_1 визнані абонентом пертинентними, а документи з K_2 - непертинентними, і по опису центра кластера $d_s \in \mathfrak{A}$ ($\mathfrak{A}$ - класифікаційна структура БД) вилічити предикати $d_s \in \mathfrak{K}_1$ і $d_s \in \mathfrak{K}_2$, де $\mathfrak{K}_1$ і $\mathfrak{K}_2$ відповідно класи пертинентних і непертинентних документів.

Нехай $\mathfrak{A}_1 = H_K(\mathfrak{A})$ - результат застосування розпізнаючого алгоритму (по типу вирахування оцінок) H до $\mathfrak{A}$ на базі навчаючої вибірки K; θ - дискретний час відносно деякої точки відліку, який пов'язаний з номером сеансу зв'язку абонента з системою (так, якщо T_0-й сеанс прийняти за точку відліку, то в сеансі $T_0 + k$: $\theta = k$ ($k \geqslant 0$)); $\mathfrak{A}^{T_0}_{\theta,z}$ - термінальні вершини структури $\mathfrak{A}_1$, знайденої в кінці сеансу θ абонента z; $[0, T^*]$ - деякий цілочисловий інтервал (T^* - деяка постійна, яка характеризує час стаціонарності інформаційної потреби абонента z) і нехай кожній вершині $a_i \in \mathfrak{A}$ певним чином поставлено у відповідність цілочисловий параметр $T_{a_i} \in [0,T^*]$, причому $\forall\, a_i \in \mathfrak{A}^{T_0}_{\theta,z}$: $T_{a_i} = T^*$. Для користувача z наступним чином побудуємо множину $A^{T_0}_{\theta,z}$:

1) при $\theta = 1$ $\quad A^{T_0}_{\theta,z} = \begin{cases} \mathfrak{A}, & T_0 = 0, \\ \emptyset, & T_0 > 0; \end{cases}$

2) при $\theta = 2$ $\quad A^{T_0}_{\theta,z} = \mathfrak{A}^{T_0}_{\theta-1,z}$;

3) при $\theta > 2$ $\quad A^{T_0}_{\theta,z} = \mathfrak{A}^{T_0}_{\theta-1,z} \cup (X_\theta - Y_\theta - W_\theta)$,

де $X_\theta = \left\{ a_i \mid a_i \in A^{T_0}_{\theta-1,z} \setminus \mathfrak{A}^{T_0}_{\theta-1,z} \right\}$, причому

$\forall\, a_i \in X_\theta$: $T_{a_i} := T_{a_i} - 1$;

$Y_\theta = \left\{ a_i \mid a_i \in X_\theta \;\&\; T_{a_i} = 0 \right\}$;

$W_\theta = \left\{ a_i \mid a_i \in X_\theta \;\&\; T_{a_i} \neq 0 \;\&\; \exists\, a_j \in \mathfrak{A}^{T_0}_{\theta-1,z} : W^{a_j}_{a_i} \right\}$,

$W^{a_j}_{a_i}$ - шлях з a_i в a_j.

Визначення 2. Множину $A^{T_0}_{\theta,z}$ будемо називати множиною активних вершин.

Змістовно, $A^{T_o}_{\theta,z}$ формується на базі зостосування алгоритма Н до $\mathfrak{A}$ з урахуванням досвіду, накопиченого за сеанс θ-1. При цьому в $A^{T_o}_{\theta,z}$ включаються всі вершини з $\mathfrak{A}^{T_o}_{\theta-1,z}$ і ті вершини з $A^{T_o}_{\theta-1,z}$ (що не увійшли в $\mathfrak{A}^{T_o}_{\theta-1,z}$), у яких значення параметрів $T_{a_1} > 0$ за умовою, що не існує шлях з a_1 в $\mathfrak{A}^{T_o}_{\theta-1,z}$. Ті вершини, котрі за час T^* не розпізнаються Н як релевантні, вилучаються з множини активних вершин.

Визначимо на множині $A^{T_o}_{\theta,z}$ лінійний порядок ρ так, щоб в першу чергу виділити найбільш актуальні вершини (на базі значень T_a), а серед них найбільш пертинентні (з використанням Н), які становлять кластери нижніх рівней ієрархічної структури $\mathfrak{A}$.

Визначення 3. Структуру $G^{T_o}_{T^*,z} = \langle A^{T_o}_{T^*,z}, \rho \rangle$ будемо називати глобальною інформаційною структурою абонента z в інтервалі часу $[T_o, T_o + T^*]$, а T^* - часом стаціонарності.

Зафіксуємо деякий інтервал часу $[T_1, T_2] \subset [T_o, T_o + T^*]$, і нехай T_1 - нова точка відліку.

Визначення 4. Структуру $L^{T_1}_{T_2,z} = \langle A^{T_1}_{T_2-T_1,z}, \rho \rangle$ назвемо локальною інформаційною структурою абонента z в $[T_1, T_2]$, якщо

$$\operatorname{card}\left\{ \bigcap_{\theta = 1,\ldots,T_2 - T_1} \mathfrak{A}^{T_1}_{\theta,z} \right\} \Big/ \max_{\theta = 1,\ldots,T_2 - T_1} \left\{ \operatorname{card}\left(\mathfrak{A}^{T_1}_{\theta,z}\right) \right\} > \mathfrak{T}, \quad (1)$$

де $\mathfrak{T}$ - деяка порогова величина.

Визначення 4. Ядром структури $L^{T_1}_{T_2,z}$ назвемо множину

$$\operatorname{Ker}\left(L^{T_1}_{T_2,z}\right) = \bigcap_{\theta = 1,\ldots,T_2 - T_1} \mathfrak{A}^{T_1}_{\theta,z} .$$

Зауваження. $\forall\, a_1 \in \operatorname{Ker}\left(L^{T_1}_{T_2,z}\right) : T_{a_1} = T^*$.

Визначення 5.

1. Якщо $\forall\, \theta \in [1, T^*] : A^{T_o}_{\theta-1,z} \cap A^{T_o}_{\theta,z} = A^{T_o}_{\theta,z}$, то будемо вважати, що абонент z має постійну інформаційну потребу. Клас таких абонентів позначимо G_1.

2. Якщо $\forall\, t \in [0, T^*] : A^{T_o}_{T^*,z} = A^{T_o}_{T^*+t,z} \,\&\, A^{T_o}_{T^*,z} \neq \mathfrak{A}$, (2) то будемо вважати, що інформаційна потреба абонента обмежена деяким постійним колом інтересів. Клас таких абонентів позначимо

через C_2.

3. Якщо не виконується (2), то будемо вважати, що коло інтересів абонента z не має фіксованих границь в заданому інтервалі стаціонарності. Клас таких абонентів позначимо C_3.

Виявимо співвідношення між глобальною і локальною структурами для різних класів абонентів.

Теорема 2. Для $z \in C_1$: $L^{T_1}_{T_2,z} = G^{T}_{T^*,z}$ з точністю до ρ.

Теорема 3. Для $z \in C_2$: $G^{T+T^*}_{T^*,z} = G^{T}_{T^*,z}$ з точністю до ρ.

Результат, отриманий в теоремі 3, дозволяє для абонента $z \in C_2$ не вести з моменту $T_o + T^*$ на протязі кванту T^* його глобальну інформаційну структуру, що значно зменшує кількість обчислень.

Нехай, як і раніше, $z \in C_2$; $\mathfrak{T}$ - деякий поріг. Розглянемо послідовність $\Omega^{T_o}_{T^*,z} = (\mathfrak{A}^{T_o}_{1,z}, \mathfrak{A}^{T_o+1}_{1,z}, \ldots, \mathfrak{A}^{T_o+T^*}_{1,z})$.

Визначення 6. Інформаційна потреба абонента z має марківський характер, якщо $\forall\, \theta \in [0,T^*]$: $\exists\, L^{T_o+\theta}_{T_o+1+\theta,z}$ & $\exists\, L^{T_o+\theta}_{T_o+\theta+2,z}$.

Очевидно, що для абонента z з марківським характером інформаційної потреби $\Omega^{T_o}_{T^*,z}$ має властивість:

$$\frac{\operatorname{card}\left\{\operatorname{Ker}\left(L^{T_o+\theta}_{T_o+\theta+1,z}\right)\right\}}{\max\left\{\operatorname{card}\left\{\mathfrak{A}^{T_o+\theta}_{1,z}\right\}, \operatorname{card}\left\{\mathfrak{A}^{T_o+\theta+1}_{1,z}\right\}\right\}} > \mathfrak{T} \;\&\; \frac{\operatorname{card}\left\{\operatorname{Ker}\left(L^{T_o+\theta}_{T_o+\theta+2,z}\right)\right\}}{\max\limits_{\theta=1,2,3}\left\{\operatorname{card}\left\{\mathfrak{A}^{T_o+\theta}_{1,z}\right\}\right\}} < \mathfrak{T}.$$

Побудуємо ймовірнісний простір $\langle\{\xi_1\}, 2^{(\xi_1)}, P\rangle$, де ξ_o - елементарна подія, яка полягає у тому, що перетин двох сусідніх елементів $\Omega^{T_o}_{T^*,z}$ є пустим; ξ_i ($i = 1,\ldots T^*-1$) - елементарна подія, яка полягає у тому, що перетин тільки $i+1$ безпосередньо слідуючих один за одним елементів $\Omega^{T_o}_{T^*,z}$ не є пустим і задоволь-

ння (1); p_i – ймовірність події ξ_i; $P : \{\xi_i\} \Rightarrow \{p_i\}$ – функція така, що $P(\xi_i) = p_i$.

Теорема 4. Якщо $z \in C_1$, то $P(\xi_{T^*-1}) = 1$. (3)

Визначення 7. Абонент $z \in C_1$ ("ξ" – слабо належить), якщо

$$\mathrm{card}\left\{\mathrm{Ker}\left(L^{T}_{T^{\circ},z}\right)\right\} \Big/ \max_{\theta=1,\ldots,T^*}\left\{\mathrm{card}\left(\mathfrak{A}^{T^{\circ}}_{\theta,z}\right)\right\} > \mathfrak{C} \quad (\mathfrak{C} \in (0,1)).$$

Теорема 5. Справедливо, що $z \in C_1 \Longrightarrow z \xi C_1$.

Теорема 6. Справедливо, що $P(\xi_{T^*-1}) = 1 \Longrightarrow z \xi C_1$.

Нехай ζ – ймовірна величина така, що $\zeta(\xi_i) = 1+i$ з ймовірністю p_i, а $M\zeta$ – її математичне сподівання. Тоді:

якщо $P(\xi_{T^*-1}) = 1$, то з урахуванням теореми 6 $z \xi C_1$ і для z на протязі часу $[T_o+T^*, T_o+2T^*]$ достатньо вести в кожний момент $\theta \in [1, T^*]$ тільки структуру $L^{T_o+T^*+\theta}_{T_o+T^*+\theta+1,z}$;

якщо $\forall [T_1, T_2] \subseteq [T_o, T_o+T^*]$: $L^{T_1}_{T_2,z} = G^{T}_{T^{\circ},z}$, то $z \in C_1$ і в цьому випадку в інтервалі часу $[T_o+T^*, T_o+2T]$ для обслуговування достатньо зберігати тільки $L^{T}_{T^{\circ},z}$;

якщо $z \in C_2$ (у відповідності з (2)), то з урахуванням теореми 3 в інтервалі $[T_o+T^*, T_o+2T]$ необхідно зберігати $G^{T}_{T^{\circ},z}$; крім цього для z з марківським характером інформаційної потреби в кожний момент $\theta \in [1, T^*]$ необхідно зберігати тільки $L^{T_o+T^*+\theta-1}_{T_o+T^*+\theta,z}$, а для z з немарківським характером інформаційної потреби в інтервалі $[T_o+T^*, T_o+2T]$ необхідно зберігати $L^{T_o+T^*+\theta}_{T_o+T^*+\theta+M\zeta}$, де $\theta = 1, 1+M\zeta, \ldots, 1+kM\zeta$; $k = [(T^*-1)/M\zeta]$.

Випадок, коли $z \in C_3$, можна інтерпретувати як ситуацію $z \in C_2$ з марківським характером інформаційної потреби при $T^* \to \infty$. Для таких z вести $G^{T}_{T^{\circ},z}$ не має сенсу, а його положення в $\mathfrak{A}$ в кожний момент θ слід визначати на базі $L^{T_o+T^*+\theta-1}_{T_o+T^*+\theta,z}$.

Таким чином, при допущені існування Н, вирішено задачу локалізації міжсеансової інформаційної потреби абонентів. (Для побудови Н можуть бути застосовані алгоритмічні схеми, які базуються на алгебраїчному підході до теорії розпізнавання образів

по Ю.І.Журавльову.)

Далі описується модель внутрішньосеансової інформаційної потреби, яка базується на гіпотезі зв'язності (ван Рійсберген, С.Джоунс): якщо термін x_1 добре розділяє класи релевантних і нерелевантних документів, то будь-який термін, що часто зустрічається разом з x_1, також володіє цією властивістю. Тому задача полягає в знаходженні множини термінів UD, які розділяють документи БД на пертинентні і непертинентні (для конкретного абонента). Крім цього, так як терміни з UD використовуються у процедурі автоматичного формулювання запиту, то бажано в UD включати як терміни, що підвищують точність пошуку, так і термини, що зменшують пошуковий шум.

Визначення 8А. Внутрішньосеансовою інформаційною потребою користувача UD^+, яка забезпечує підвищення точності пошуку, будемо називати множину термінів БД, ймовірність яких знаходитись в пертинентних документах вище деякого порогу M^+.

Визначення 8В. Внутрішньосеансовою інформаційною потребою користувача UD^-, яка забезпечує мінімізацію пошукового шуму, будемо називати множину термінів БД, ймовірність яких знаходитись в непертинентних документах вище деякого порогу M^-.

Нехай $D_1=(d_i)$ - множина документів, переглянутих абонентом на протязі першого циклу роботи процедури ранжування, і нехай $T_1=(t_j)$ - множина термінів, які знаходяться у всіх цих документах. Розглянемо трійку $\langle t_j,a_j,b_j\rangle$, де $t_j\in T_1$, a_j - кількість документів, які відмічені абонентом як пертинентні і в яких знаходиться термін t_j, b_j - кількість документів в БД, в яких знаходиться термін t_j.

На множині $X_1=(\langle t_j,a_j,b_j\rangle)$ визначимо функцію I: $X_1\to\mathbb{R}$ (де $\mathbb{R}$ - множина дійсних чисел), яку будемо називати інформативністю терміна. Як приклад може виступати

$$I(t_j)=\log\left[(p_j\cdot(1-q_j))/(q_j\cdot(1-p_j))\right],$$

де $p_j=\dfrac{a_j+0.5}{N+1}$, $q_j=\dfrac{(b_j-a_j)+0.5}{M+1}$, N - загальна кількість пертинентних документів, M - кількість документів БД за виключенням N.

Змістовно, інформативність відображає здатність терміну розділяти документи БД на пертинентні і непертинентні.

Видно, що задача локалізації внутрішньосеансової інформа-

ційної потреби може бути зведена до задачі виділення з множини термінів, які входять в уже переглянуті користувачем документи, таких термінів, у яких інформативність більше (для позитивних значень інформативності) або меньше (для негативних значень інформативності) деякого порогу.

Розглянемо задачу визначення UD^+. Нехай $I^+ = (I_i^+)$ - множина різних позитивних значень інформативності термінів T_i. Побудуємо ймовірнісний простір $\langle\{\xi_i\}, 2^{\{\xi_i\}}, P\rangle$, де ξ_i $(i=1,\ldots,card(I^+))$ - елементарна подія, яка полягає у тому, що інформативність терміна прийме значення I_i^+; p_i - ймовірність події ξ_i; $P: \{\xi_i\} \Rightarrow (p_i)$ - функція така, що $P(\xi_i) = p_i$.

Нехай ζ - випадкова величина така, що $\zeta(\xi_i) = I_i^+$ з ймовірністю p_i, а $M\zeta$ - її математичне сподівання.

Тоді з T_i виділимо підмножину термінів $\mathfrak{T}_i^+$ таку, що для будь-якого $t_j \in \mathfrak{T}_i^+$ інформативність I_j терміну t_j більша або дорівнює $M\zeta$. З X_i виділимо підмножину $\mathfrak{X}_i^+$, яка відповідає $\mathfrak{T}_i^+$. Побудована таким чином $\mathfrak{X}_i^+$ (якщо покласти $M^+ = M\zeta$) якраз і є внутрішньосеансовою інформаційною потребою користувача, яка забезпечує підвищення точності інформаційного пошуку перед другим циклом роботи процедури ранжування.

Для i-ої $(i \geqslant 2)$ пошукової транзакції T_i будується наступним чином: $T_i = \mathfrak{T}_{i-1}^+ \cup T(D_i)$, де $T(D_i)$ - множина термінів документів D_i, які були переглянуті користувачем за час i-ої транзакції. Для T_i будується відповідна їй множина $\mathfrak{X}_i^+$, яка в свою чергу є внутрішньосеансовою інформаційною потребою користувача, що забезпечує підвищення точності інформаційного пошуку перед наступною пошуковою транзакцією.

Ймовірність p_i випадкової події ξ_i будемо відновлювати таким чином, щоб більші значення інформативності мали більшу ймовірність. Тоді p_i і $M\zeta$ можна обчислювати за формулами:

$$p_i = I_k^+ \Big/ \sum_{I_k^+ \in I^+} I_k^+ \; ; \quad M\xi = \left(\sum_{I_k^+ \in I^+} \left[I_k^+ \right]^2 \right) \Big/ \sum_{I_k^+ \in I^+} I_k^+ .$$

Множина $\mathfrak{X}_i^+$ може бути використана для підвищення точності пошуку процедурою формування запиту на (i+1)-ому циклі роботи у схемі адаптивного інформаційного пошуку.

Для визначення UD^- слід ураховувати терміни з негативними

значеннями інформативності. У цьому випадку їх можна було б використовувати в запитах зі зв'язкою "не". Для здобуття таких термінів, аналогічно $\mathfrak{T}_i^+$, будується множина $\mathfrak{T}_i^-$, яка містить терміни t_i, у яких значення інформативності негативне і менше M_{ζ^-}, де ζ^- - випадкова величина, визначена на множині різних негативних значень інформативностей.

В залежності від того, до якого класу буде віднесено конкретного користувача, застосовуються наступні дії:

1. Для користувачів з постійною інформаційною потребою термінологічні словники не змінюються і використовуються для автоматичного забезпечення режиму вибіркового розповсюдження інформації.

2. Для користувачів, інформаційна потреба у яких змінюється в межах деякого постійного кола інтересів, процедура локалізації внутрішньосеансової інформаційної потреби постійно провадить "рафінування" термінологічного словника перед кожною новою пошуковою транзакцією.

3. Для користувачів, у яких інформаційна потреба постійно змінюється, терминологічні словники не зберігаються і такі абоненти повинні переглядати кластерне дерево (для початкового позиціонування в БД) перед черговим сеансом зв'язку з системою.

Друга глава присвячена опису семантичної моделі та її використанню у схемі адаптивного інформаційного пошуку. Ключові моменти пропонованої моделі полягають у наступному.

Нехай $X = \{x_i\}$ - множина термінів, які позначають поняття $P = \{p_i\}$ вихідної предметної галузі, а $\hat{\Omega}$- множина парадигматичних відношень на множині P (всі парадигматичні відношення є толерантними або транзитивно-антисиметричними).

Відношення $\hat{\omega} \in \hat{\Omega}$ індукує відношення ω на множині X:

$$\forall\, p_i,\ p_j \in P:\ p_i \hat{\omega}\ p_j \Longrightarrow x_i \omega\ x_j.$$

Транзитивне відношення $\omega \in \Omega$ ε-(ε^{-1})-орієнтоване, якщо $\forall\ (x_i, x_j) \in \omega : M_i \rho\ M_j$ ($M_i \rho^{-1} M_j$), де M_i, M_j - множина об'єктів, які відповідають поняттям p_i і p_j, ρ - деяке відношення уточнення властивостей, дроблення об'єктів і т.п. Додамо до Ω відношення зворотне до транзитивних, розглянемо транзитивне замкнення Ω^+ множини Ω і введемо на Ω^+ асоціативну операцію "*" добутку відношень: $x_i \omega_1 * \omega_2 x_j \Longleftrightarrow \exists\, x_s :\ x_i \omega_1 x_s\ \&\ x_s \omega_2 x_j$ ($\omega_1, \omega_2 \in \Omega^+$). Замикаючи Ω^+ відносно "*", утворимо універсальну алгебру складних від-

ношень $U_{\Omega^c} = \langle \Omega^c, (*) \rangle$.

Представлення $\omega \in \Omega^c$ у вигляді $K(\omega) = \omega_{i_1} * \ldots * \omega_{i_s}$ назвемо канонічним, якщо $\forall\, k = 1, \ldots, s-1 : \omega_{i_k} * \omega_{i_{k+1}} \neg\in \Omega^+$.

Теорема 1. Для будь-якого $\omega \in \Omega^c$ $K(\omega)$ єдине.

Добуток $\omega_1 * \omega_2$ назвемо добутком $\omega_{1k}\omega_{21}$-типу, якщо

$$K(\omega_1) = \omega_{11} * \ldots * \omega_{1k} \quad \text{і} \quad K(\omega_2) = \omega_{21} * \ldots * \omega_{2s}.$$

Визначення 1. Канонічне представлення називається направленим, якщо в ньому не зустрічаються добутки $\omega_{1j}\omega_{1j}^{-1}$-типу.

Теорема 2. Якщо $\omega_1 * \omega_2$ добуток $\omega_{1k}\omega_{21}$-типу і $\omega_{1k} \neq \omega_{21}$, то $K(\omega_1 * \omega_2) = K(\omega_1) * K(\omega_2)$.

Наслідок. В канонічному представлені не можуть зустрічатися добутки $\omega_i\omega_i$-типів.

Можливість представлення складних відношень у канонічному вигляді дозволяє виконувати кількісну оцінку складних відношень на базі простих.

Використовуючи властивості алгебри U_{Ω^c}, будується функція семантичного зв'язку f: $\Omega^c \to (0,1)$. При цьому враховуються як топологічні характеристики графа залежності (довжина шляху між термінами, ширина основи поділу, тип відношень відповідних компонентів зв'язності), так і деякий набір семантичних передумов (характерних для задач кластеризації БД, локалізації інформаційної потреби, інформаційного пошуку), які обумовлюють властивості функції f (симетричність; значення f для транзитивних відношень більше ІІ значень для толерантних відношень; значення f для відношень ω, канонічні представлення яких не містить толерантних відношень, перевершують значення f для відношень, які містять толерантні відношення; значення f узгоджені з порядком τ, який задається на Ω^+).

Задамо на Ω лінійний порядок τ і перенумеруємо відношення у відповідності з цим порядком: $\omega_i \tau\, \omega_j \iff i < j$. Тоді для складного відношення ω, з урахуванням вигляду графа залежності і введених семантичних передумов, функція семантичного зв'язку має вигляд:

$$f_{\omega_{j_1}\ldots\omega_{j_m}\omega_{i_1}\ldots\omega_{i_n}}(x,y) = \prod_{l=1}^{m}\left(A_{j_l}\right)^{k_l} \prod_{k=1}^{n} \frac{1}{B_{i_k}} \left[1 - \sum_{k=1}^{n}\left(\frac{B_{i_k}}{B_{i_{k-1}}}\right)^{k} \sum_{s=1}^{p} \tau_{q_s}^{OT}\right],$$

де B_{i_k} - параметр, який відповідає транзитивному відношенню ω_{i_k};

A_{j_1} - параметр, який відповідає толерантному відношенню ω_{j_1}; p - довжина шляху i_k-ої транзитивної компоненти складного відношення; k_1 - кількість ланок толерантного відношення ω_{j_1}, які входять у шлях, що з'єднує терміни x і y; n - кількість транзитивних відношень в $K(\omega)$; m - кількість толерантних відношень в $K(\omega)$; $\gamma_{\mathfrak{A}_s}^{OT}$ - відносний коефіцієнт уточнення, який є функцією від основи поділу; параметри B_{i_k} і A_{j_1} задовольняють умовам

$$B_{i_o} \equiv B_{i_1}; \quad 0 < A_{j_1} < A_{j_{1+1}} < 1; \quad B_{i_k} > B_{i_{k+1}} > 1.$$

Для транзитивних (простих і складних) відношень значення функції f_ω повинні задовольняти умові

$$L_1 < f_{\omega_{i_1} \ldots \omega_{i_n}} < L_2, \qquad (1)$$

де

$$L_1 = \begin{cases} \max f_{\omega_{i_1} \ldots \omega_{i_j - 1}}, & \text{якщо } i_n = N, \text{ де } j \in \{1, \ldots, n\} \text{ таке,} \\ & \text{що } \forall k = \overline{0, n-j}\ \ i_{j+k} = i_j + k; \\ \max f_{\omega_{i_1} \ldots \omega_{i_n} \omega_N}, & \text{у іншому випадку;} \end{cases}$$

$$L_2 = \begin{cases} \min f_{\omega_{i_1} \ldots \omega_{i_{n-1}}}, & \text{якщо } i_N = N; \\ \min f_{\omega_{i_1} \ldots \omega_{i_{n-1}} \omega_{i_n+1} \omega_{i_n+2} \ldots \omega_N}, & \text{у іншому випадку;} \end{cases}$$

N - кількість транзитивних відношень; $i_o = N+1$; $f_{\omega_{N+1}} \equiv 1$ і $f_{\omega_o} \equiv 0$.

Для транзитивно-толерантних відношень, які містять тільки по одному простому толерантному відношенню, значення функції f_ω повинно задовольняти умові ($k_m = 1,2$):

$$\left(\left(A_{j_m} \right)^{k_m} \cdot L_1, \left(A_{j_m} \right)^{k_m} \cdot L_2 \right), \qquad (2)$$

Теорема 3. Для того, щоб $0 < f_\omega < 1$, достатньо

$$B_i / B_j \leqslant 2 / \left[1 + \sqrt{1 + 4\gamma_{\mathfrak{A}}^{max}} \right], \quad i > j; \quad 0 < A_k < 1;$$

де $\gamma_{\mathfrak{A}}^{max} = \max(\gamma_{\mathfrak{A}}) \cdot \max(\max(r_{ik}))$, r_{ik} - довжина k-го шляху в i-ій компоненті зв'язності, $\gamma_{\mathfrak{A}}$ - ширина основи поділу, $\mathfrak{A}$ - клас основи поділу.

Теорема 4. Рекурентні співвідношення

$$B_{N-1} \geqslant \gamma_{\mathfrak{A}}^{max} \cdot B_N / b; \qquad B_i \geqslant \gamma_{\mathfrak{A}}^{max} \cdot B_{i+1} \cdot \ldots \cdot B_N / b;$$

де $0 < b < 1$, є достатніми для виконання (1).

Наслідок 1. Значення параметрів B_i, $1=\overline{1,N}$, можна обчислювати за формулою $B_i=\left(\gamma_{\mathfrak{A}}^{max} \Big/ b\right)^{2^{N-1}}$.

Наслідок 2. Умови

$A_i<A_j$; $A_i>A_M^2$; $A_M\le\left[1-\gamma_{\mathfrak{A}}^{max}\cdot\frac{c^{N-1}-1}{c^N(c-1)}\right]\Big/\left[(\gamma_{\mathfrak{A}}^{max}+1)\cdot c^{2^{N-1}}\right]=L$, $A_j<A_{j+1}\cdot L$;

де $i<j$, M – кількість толерантних відношень в Ω, $c=\gamma_{\mathfrak{A}}^{max}/b$, є достатніми для виконання (2).

Наслідок 3. Лінійний порядок, індукований по f_ω на Ω природним порядком на $(0,1)$, співпадає з τ.

Побудуємо скалярно-семантичну і семантичну функції корреляції документів d_i, d_j.

1. Скалярно-семантична функція корреляції документів f^* вибирається у вигляді $f^* \stackrel{df}{=} S\left(\sum_{t\in d_j}\phi_t(d_j)\right)$, де функція ϕ_t визначається наступним чином (t – термін):

1) $\forall\ t:\ \varphi_t(d_i)\cdot\varphi_t(d_j)=1 \Longrightarrow \psi_t(d_j)=\varphi_t(d_j)$;

2) $\forall\ t:\ \varphi_t(d_i)\cdot\varphi_t(d_j)=0 \Longrightarrow \psi_t(d_j)=\max\limits_{r\in d_j;\ \omega\in\Omega^\circ}\{f_\omega(t,r)\}$

і $\varphi_t(d_i)=1$, якщо $t\in d_i$, і $\varphi_t(d_i)=0$ – у протилежному випадку.

2. Скалярно-семантична функція f^* дозволяє оцінити лише найбільш суттєві семантичні зв'язки між термінами порівнюваних документів. Для врахування всіх семантичних зв'зків визначаються поняття семантичної сили терміну $\beta_{i,K}$ у деякій множині термінів $K=\{t_{i_1},\dots,t_{i_s}\}$ і узагальнююче його поняття семантичної сили множини термінів K_1 у деякій множині термінів K_2 – β_{K_1,K_2}:

1) $\beta_{i,K}=\sum\limits_{t_{i_j}\in T_i\cap K}\sqrt[n]{f_\omega(1,1_j)}$, $T_i=\{t_{i_j}\mid \exists\,\omega\in\Omega^\circ:\ t_i\omega\, t_{i_j}\}$,

причому якщо існує декілька $\tilde\omega\in\Omega^\circ$: $t_i\tilde\omega\, t_{i_j}$, то для визначення f_ω вибирається відношення, яке дає найбільшу силу зв'язку, тобто $f_\omega(1,1_j)=\max\limits_{\tilde\omega\in\Omega^\circ}\{f_{\tilde\omega}(1,1_j)\}$;

2) $\beta_{K_1,K_2}=\sum\limits_{t_i\in K_1}\beta_{i,K_2}$.

Тоді семантичну функцію f^{**} корреляції документів, яка вра-

ховує всі взаємозв'язки досліджуємих документів, можна визначити наступним чином: $f^{**}(d_i,d_j) = \rho_{d_i,d_j} / \min(\rho_{d_i,T},\rho_{d_j,T})$.

Відмітимо, що функції f^{*} і f^{**} узагальнюють кореляційні функції, у яких враховується тільки синонімія.

Третя глава присвячена опису реалізованого автором програмного прототипу системи адаптивного інформаційного пошуку та аналізу проведених експериментів. Даний прототип реалізовано в оточенні IPIC, усі програмні компоненти написані на мові Сі. Для роботи системи необхідно мати персональний комп'ютер, сумісний з IBM PC/AT у стандартній конфігурації, та операційну систему MS DOS версії 3.0 та вище. Для експериментів використовувалась БД в галузі обчислювальної техніки і програмного забезпечення (300 документів). При реалізації системи велика увага приділялася інтерфейсу користувача, який був зведений до того, що від абонента вимагалося лише здійснювати оцінку пред'явлених йому системою документів. Вся решта роботи по визначенню інформаційної потреби, формулюванню запиту, здійсненню пошукових ітерацій, упорядкуванню результатів пошуку перед переглядом з урахуванням інформаційних інтересів користувача здійснювалась автоматично самою системою. Для кожного абонента система створює і веде його персональну віртуальну БД.

Експерименти, проведені на даній системі, показали, що вона досить добре адаптується до інформаційної потреби користувача. Середній показник точності при майже 100 % повноті становив біля 50 % (у звичайних пошукових системах показник точності не перевершує 20%), характер оцінок документів у пред'являємих системою порціях мав коливальний вигляд, тобто погано - краще - добре - погано - краще - добре - ... Причому перехід до слідуючого періоду означає, що система вже вичерпала розділяючі властивості найбільш інформативних термінів і намагається "зачепитися" за іншу лексику, семантично близьку до вибраної тематики.

У таблиці приведено оцінки документів, які давались у кожній порції, пред'являємої системою, та Стоп-правила, які використовувала система у випадку адаптації II на інформаційні інтереси у галузі мережевого забезпечення.

Після всіх шагів "рафінування" словника користувача він містив 124 терміна, які повністю покривали мережеву термінологію

БД (відмітимо, що на початковій стадії, після проглядання кластерного дерева, словник містив 11 термінів).

	Порція	Оцінки документів										Стоп-правило
		d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	
$\varphi(s_0)\neq\emptyset$	1	+	+	-	-	-	-	-				переупорядкувати
	2	+	-	+	+	+	+	-	-	+	-	переупорядкувати
	3	+	+	+	-	+	+	-	-	-	+	переупорядкувати
	4	+	+	+	+	+	+	+	-	+	+	продовжити
	5	+	+	+	+	+	+	+	-	-	-	продовжити
	6	-	-									новий запит
$\varphi(s_1)\neq\emptyset$	1	-	+									новий запит
$\varphi(s_2)\neq\emptyset$	1	+	+	-								новий запит
$\varphi(s_3)=\emptyset$												КІНЕЦЬ

"+" - документ відмічений як релевантний,

"-" - документ відмічений як нерелевантний.

Таблиця.

Основні виводи і результати, отримані автором по темі дисертації, полягають у слідуючому:

1. Автором запропонована нова схема інформаційного пошуку - адаптивний інформаційний пошук,- яка є розвитком ітераційних пошукових схем.

2. Для реалізації цієї схеми автором запропонована модель локалізації міжсеансової інформаційної потреби абонента. Запропоновані стратегії локалізації міжсеансової інформаційної потреби, вивчені види її мінливості. З цією метою введені поняття глобальної і локальної інформаційних структур, виділено три класи абонентів і знайдено умови віднесення абонента до того чи іншого класу на підставі співвідношень між цими структурами.

3. Для локалізації інформаційної потреби в межах одного сеансу зв'язку абонента з системою введено поняття внутрішньосеансової інформаційної потреби і запропоновано ймовірнісний підхід до побудови відповідних термінологічних словників.

4. Запропонована семантична модель інформаційно-пошукових процесів для задач кластеризації БД, локалізації інформаційної потреби користувача, пошуку і ранжування найдених документів. Побудована всюди визначена функція сили зв'язку між двома термінами, на базі якої побудовані скалярно-семантична і семантична

функції кореляції документів.

5. Для апробації схеми адаптивного інформаційного пошуку розроблено і реалізовано прототип системи адаптивного пошуку. При реалізації системи велика увага приділялась інтерфейсу користувача, який був зведений до того, що від абонента вимагалось лише здійснювати оцінку пред'явлених йому системою документів. Вся решта роботи по визначенню інформаційної потреби, формулюванню запиту, здійсненню пошукових ітерацій, упорядкуванню результатів пошуку перед переглядом з урахуванням інформаційних інтересів користувача здійснювалась автоматично самою системою.

6. З метою перевірки адекватності моделей була проведена серія експериментів з використанням експериментальної БД в галузі обчислювальної техніки і програмування. Експерименти показали, що дана схема і використані автором для її реалізації моделі дають, по-перше, добрі показники повноти і точності інформаційного пошуку і, по-друге, термінологічний словник користувача, який будується системою, фактично покриває всю представницьку (природньо, у статистичному розумінні) термінологію по вибраній темі в конкретній БД.

Основні результати дисертації опубліковані в роботах:

1. Дриянский В.М., Жогов Г.В. Модель локализации информационной потребности абонентов настраивающихся документальных информационно-поисковых систем//Кибернетика. -1986. - N 4. - С. 81-84.

2. Дриянский В.М., Жогов Г.В. Система программных интерфейсов // Базы данных и знаний в автоматизированных региональных системах. - К.: Наукова думка, 1990. - С. 269 - 278.

3. Дриянский В.М., Жогов Г.В., Алешкина С.М. и др. Персональная электронная библиотека // Проблемы информатики (Научно практическая конф. с межд. участием). - Самара-Астрахань-Самара, 11 - 18 мая 1991. - С. 35 - 37.

4. Дриянский В.М., Жогов Г.В., Катеринич Л.Г. Алгоритм принятия решения в моделях локализации входной информационной потребности // Проблемы создания ретроспективных поисковых массивов в автомат. центрах НТИ. Тез. док. XV Всес. научн. сем. "Систем. исследов. ГАСНТИ" (г.Рига, 20 - 22 мая 1985) Ч.I - С. 134-135.

5. Дриянский В.М., Жогов Г.В., Катеринич Л.Г. Алгоритмы классификации в моделях локализации инф. потребности пользовате-

лей документальной информационно-поисковой системы // Математические методы в автоматизированных информационных системах и банках данных. - К.: ИК АН Украины, 1985. - 49 - 55.

6. Дриянский В.М., Жогов Г.В., Катеринич Л.Г. Концепция настраивающегося информационного поиска в АСИО по ВТ // Создание автоматизированных систем информационного обеспечения научных исследований. - К.: ИК АН Украины, 1986. - С. 27 - 38.

7. Дриянский В.М., Жогов Г.В., Катеринич Л.Г. Математические модели настраивающегося документального поиска//В кн. Автоматизация информ. поиска.- К.:Наукова думка,1990. - С. 121 - 181.

8. Дриянский В.М., Жогов Г.В., Катеринич Л.Г. Нужен ли язык запросов в информационно-поисковых системах // УСиМ. - 1991. - N 7. - С. 158 - 160.

9. Дриянский В.М., Жогов Г.В., Капустин В. А., Катеринич Л.Г. Информационно-поисковая система с открытой архитектурой // Проблемы информатики города: Сб. науч. тр. - К.: Наукова думка, 1990. - С. 218 - 228.

10. Дриянский В.М., Жогов Г.В., Колтун А.Я. Количественные оценки парадигматических отношений и их использование в документальных ИПС // НТИ. Сер. 2. - 1985. - N 10. - С. 6 - 16.

11. Дриянский В.М., Жогов Г.В., Чалый И.А. Концепция операционной оболочки для создания информационных технологий // Проблемы разработки и внедрения программного обеспечения ЭВМ и систем. - К.: ИК АН Украины, 1988. - С. 24 - 30.

12. Дриянский В.М., Жогов Г.В., Чалый И.А. Мобильные инструментальные средства поддержки протоколов коммуникации прикладных процессов//Программирование. - 1990. - N 6. - С. 103-108.

Підписано до друку: 20 .09. 94г . Формат 60×84 1/16 .
Об'єм 1,05 д.а. Зам.N 4034 . Тираж 100 примірників.

Державне комунальне поліграфічне підприємство "Тираж"
м.Київ

Ав 31.003