УДК 681.3

О.Ю. СЕРГЄЄВ[1], С.В. СОМОВ[1], І.Ю. СУБАЧ[2]

[1]*Полтавський військовий інститут зв'язку, Україна*
[2]*Військовий інститут телекомунікацій та інформатизації НТУУ "КПІ", Україна*

МЕТОД МОДИФІКАЦІЇ ВИХІДНИХ ЗАПИТІВ КОРИСТУВАЧІВ АВТОМАТИЗОВАНИМИ ІНФОРМАЦІЙНО-ПОШУКОВИМИ СИСТЕМАМИ

Пропонується новий метод модифікації вихідних запитів користувачів автоматизованими інформаційно-пошуковими системами, шляхом розширення вихідного запиту найбільш інформативними ключовими словами з масиву ключових термів релевантних документів, підготовлених автоматизованою інформаційно-пошуковою системою на цей запит.

автоматизована інформаційно-пошукова система, інформаційний пошук, модифікація інформаційного запиту, ключове слово, пошуковий образ документа

Вступ

Для знаходження необхідної електронної інформації в мережі *Internet* користувачу доводиться використовувати спеціальні пошукові інструменти. Найчастіше в ролі таких інструментів виступають пошукові системи і каталоги. Аналіз масиву запитів до пошукових систем показує, що користувач повинен неодноразово перефразовувати запит, перш ніж він отримає список знайдених документів, що задовольняють його інформаційні потреби.

При вирішенні задач інформаційного пошуку автоматизованими інформаційно-пошуковими системами (АІПС), зокрема, оцінці відповідності знайдених документів інформаційним потребам користувача, існують деякі фактори, які впливають на якість рішення цих задач, серед яких точність формулювання пошукового запиту має першочергове значення [1]. Інформаційні ресурси, посилання на які повертає АІПС у відповідь користувачу, найчастіше, взагалі не відносяться до його інформаційних потреб здебільшого тому, що саме вихідний запит був сформований невірно.

Задачу підвищення ефективності пошуку інформації за допомогою універсальних пошукових машин можна спробувати вирішити шляхом модифікації вихідного запиту користувача автоматичним додаванням слів, що у найбільшому ступені описують інформаційні ресурси, які цікавлять користувача, і рекомендуючи йому цей модифікований запит для подання в АІПС.

В даний час є спроби модифікувати запит за допомогою тезауруса, нечіткого пошуку і пошуку за зразком, онтології, зворотного зв'язку з користувачем, на основі асоціативних зв'язків та інших [2 – 4].

Технологія нечіткого пошуку дозволяє модифікувати запит близькими за написанням словами, що містяться в колекції документів, по яких ведеться пошук. Цей підхід скоріше призначений для відстеження і виправлення орфографічних помилок у запиті користувача, чим для його модифікації.

Тезаурус дозволяє модифікувати запит близькими за змістом словами, використовуючи різні типи значеннєвих зв'язків. Пошук документів за зразком дозволяє знайти документи, близькі за змістом до заданого. Як модель змісту тексту, при порівнянні документів, використовується семантична мережа чи набір ключових тем. Дані підходи так чи інакше використовують знання людини (розроблювача) АІПС про предметну область, у якій буде відбуватися пошук. Для реалізації АІПС, заснованих на даних підходах, необхідно будувати семантичні мережі і

тезауруси, що ускладнює розробку інформаційної системи і приводить до втрати релевантності інформації, яка видається, якщо яке-небудь поняття не включене в мережу чи тезаурус.

При модифікації на основі асоціативних зв'язків запит будується на основі масиву запитів, що надійшли в систему за визначений час. У цьому перевага і недолік такої методики. При побудові асоціацій вважається, що користувач визначений час тільки уточнював вихідний запит. Якщо ж користувач спочатку цікавився однією предметною областю, а потім став цікавитися іншою, то будується неправильна асоціація і це приводить до модифікації запиту невірними словами.

У зв'язку з цим, виникає актуальна задача розробки нових підходів до модифікації вихідних запитів користувачів, що були б позбавлені перерахованих вище недоліків.

У статті пропонується новий метод модифікації вихідних запитів користувачів автоматизованими інформаційно-пошуковими системами, шляхом розширення вихідного запиту найбільш інформативними ключовими словами з масиву ключових термів релевантних документів, підготовлених автоматизованою інформаційно-пошуковою системою на цей запит.

Математичний апарат

Класичні теоретико-множинні моделі інформаційного пошуку розглядають документи

$$D = \{d_1,\ d_2,\ ...,\ d_s\}$$

і запити користувачів

$$Z = \{z_1,\ z_2,\ ...,\ z_m\}$$

як множини ключових слів (термів)

$$T = \{t_1,\ t_2,\ ...,\ t_e\},$$

з яких вони складаються, де під термом $t_{e'}$, $e' = \overline{1,e}$ розуміють слово, семантика якого дозволяє описати основний зміст документа $d_{s'} \in D$, $s' = \overline{1,s}$ і запиту $z_{m'} \in Z$, $m' = \overline{1,m}$.

Для вирішення задачі модифікації первинного запиту користувача $z_{m'} \in Z$, $m' = \overline{1,m}$ необхідно визначити множину термів $K(z_{m'})$ для модифікації цього запиту. Кожна інформаційно-пошукова система оперує не з самим документом, а з його пошуковим образом (ПОД) – множиною ключових слів (термів) даного документа. Документ $d_{s'} \in D$, $s' = \overline{1,s}$ і масив ключових термів (пошуковий образ документа – ПОД) $T^{d_{s'}} = \{t_1,\ t_2,\ ...,\ t_f\}$ цього документа в інформаційно-пошукових системах – еквівалентні поняття. Тому в якості множини документів, з якої будуть відбиратися терми для модифікації запиту, доцільно використовувати масив найбільш релевантних документів $D_r \subset D$, які були відібрані на первинний запит та мають найбільш вагоме значення для потреб користувача, які виражені через його запит. Визначення даного масиву може бути проведено за допомогою методики оцінки релевантності, яку описано в [5].

Для модифікації запиту необхідно визначити, які саме ключові терми можуть розширити первинний запит. Дана задача розкладається на дві підзадачі: визначення множини документів $D_c \subset D_r$, терми з яких можуть брати участь у модифікації запиту, та визначення термів для розширення.

Рішення першої задачі ускладнюється внаслідок того, що неможливо впевнено сказати, наскільки і які саме документи відносяться до множини D_c, а які ні, до того ж чіткий поділ об'єктів на класи за обмірюваним значенням деякої ознаки є значним спрощенням предметної області. При рішенні практичних задач, бажана більш природна для фахівців оцінка ступеня приналежності об'єктів до класів – лінгвістична апроксимація. Для цього доцільно використовувати апарат нечітких множин (НМ).

Для визначення множини $D_c \subset D_r$, терми з документів якої можуть брати участь у модифікації первинного запиту, введемо лінгвістичну змінну

<Множина документів для модифікації запиту, G, $W>$, де G – терм-множина лінгвістичної змінної, яка визначається сукупністю її лінгвістичних термів $G=\{g_1,g_2,...,g_v\}$; $W=[w_1,w_n]$ – базова множина лінгвістичної змінної. Назва кожного лінгвістичного терму співпадає з назвою множини щодо модифікації запиту. Лінгвістичні терми $g_{v'}\in G$, $v'=\overline{1,v}$ розглядатимемо як нечіткі множини, задані на універсальній множині W.

Дані множини визначимо як

$$\widetilde{g}_{v'}=\int_{w_1}^{w_n}\mu_{g'}(w_i)/w_i, \qquad (1)$$

де $\mu_{g'}(w_i)$ – функція приналежності (ФП) значення $w_i\in W$, $i=\overline{1,n}$ терму $g_{v'}\in G$, $v'=\overline{1,v}$. У співвідношенні (1) знак інтегралу позначає об'єднання пар $\mu(u)/u$.

При визначенні множини документів для модифікування запиту $D_c\subset D_r$, потрібно побудувати функції приналежності $\mu_{g_{v'}}$ лінгвістичних термів $g_{v'}\in G$, $v'=\overline{1,v}$, тобто визначити значення $\mu_{g_{v'}}(w_i)$, $g_{v'}\in G$, $v'=\overline{1,v}$, $w_i\in W$, $i=\overline{1,n}$.

Функція приналежності $\mu_{g_{v'}}$ визначається за матрицею попарних порівнянь $M=\|m_{ij}\|$, елементи якої m_{ij} являють собою деякі оцінки інтенсивності приналежності елементів $w_i\in W$ нечіткій множині $\widetilde{g}_{v'}$ у порівнянні з елементами $w_j\in W$ [6]. Нехай значення функції приналежності $\mu_{g_{v'}}$ відомі для всіх елементів $w\in W$, наприклад $\mu_{g_{v'}}(w_i)=r_i$ $(i\in|W|=\{1,2,...,n\})$, тоді попарні порівняння представляються матрицею відносин M, де $m_{ij}=r_i/r_j$. За умови точності відносин одержимо співвідношення $Mr=nr$, $r=(r_1,\ r_2,\ ...,\ r_n)$, де n – власне значення матриці M, за яким можна відновити вектор r (з урахуванням умови $\sum_{i=1}^{n}r_i=1$). Із [6, 7] маємо, що в загальному випадку емпіричний вектор $r=(r_1,\ r_2,\ ...,\ r_n)$ повинний задовольняти задачі на пошук власного значення $Mr=\lambda_{\max}$, де $\lambda_{\max}$ – найбільше власне значення, і задача зводиться до пошуку вектора r, що задовольняє рівнянню $Mr=r\lambda_{\max}$. Так як відомо, що це рівняння має єдине рішення, то значення координат власного вектора, що відповідає максимальному власному значенню, поділені на їхню суму, будуть шуканими ступенями приналежності.

Матриця попарних порівнянь будується шляхом опитування експерта (експертів) щодо того, наскільки, на його думку, величина $\mu_{g_{v'}}(w_i)$, $w_i\in W$ перевищує величину $\mu_{g_{v'}}(w_j)$, $w_j\in W$, тобто наскільки документи з діапазону w_i більш значимі для терму $g_{v'}\in G$, $v'=\overline{1,v}$, ніж документи з діапазону w_j. Поняття, якими може оперувати експерт, представлені в табл. 1.

Як видно з таблиці, для поліпшення погодженості оцінок передбачається, що $m_{ij}m_{jk}=m_{ik}$, де $m_{ii}=1$ для діагональних елементів і $m_{ij}=1/m_{ji}$ для елементів, симетричних щодо головної діагоналі.

Матриця відносин M має наступний вигляд:

$$M=\begin{Vmatrix} r_1/r_1 & r_1/r_2 & \cdots & r_1/r_n \\ r_2/r_1 & r_2/r_2 & \cdots & r_2/r_n \\ \cdots & \cdots & \cdots & \cdots \\ r_n/r_1 & r_n/r_2 & \cdots & r_n/r_n \end{Vmatrix}.$$

Для визначення i-го елемента вектора r $(i\in|W|)$ обчислимо суму елементів i-го стовпця матриці M. Ця сума дорівнює деякому числу k_j, тобто $\sum_{i=1}^{n}m_{ij}=k_j$. З побудови матриці одержуємо, що

$$\sum_{i=1}^{n} m_{ij} = \sum_{i=1}^{n} \frac{r_i}{r_j} = \frac{\sum_{i=1}^{n} r_i}{r_j} = \frac{1}{r_j} \ .$$

Таким чином, $r_j = 1/k_j$. Продовжуючи процедуру по всіх стовпцях матриці M, будується шуканий вектор r, значення якого після нормалізації і будуть шуканими значеннями функції приналежності для терму $g_{v'} \in G$, $v' = \overline{1,v}$.

Таблиця 1

Поняття, якими може оперувати експерт при побудові матриці попарних порівнянь

Інтенсивність важливості	Якісна оцінка	Пояснення
0	Незрівнянність	Нема рації порівнювати елементи
1	Однакова значимість	Елементи рівні за значенням
3	Слабко значиміше	Існують показання в перевазі одного елемента іншому, але показання непереконливі
5	Істотно чи сильно значиміше	Існують гарні докази і логічні критерії, які можуть показати, що один з елементів найбільш важливий
7	Очевидно значиміше	Існують переконливі докази більшої значимості одного елемента в порівнянні з іншим
9	Абсолютно значиміше	Максимально підтверджується відчутність переваги одного елемента іншому
2, 4, 6, 8	Проміжні оцінки між сусідніми оцінками	Необхідний компроміс
Якщо оцінка m_{ij} має ненульове значення, приписане на підставі порівняння елемента r_i з елементом r_j, то m_{ji} має зворотне значення $1/m_{ij}$		

Якщо кількість термів лінгвістичної змінної дорівнює трьом та в якості базової множини виступає кількість документів у відсотках множини $D_r \subset D$, то графік ФП для цих термів набуває вигляду, зображеного на рис. 1. Для визначення множини $D_c \subset D_r$ введемо умову:

якщо $\mu_{g_1}(w_i) > \mu_{g_2}(w_i) > ... > \mu_{g_v}(w_i)$,

тоді $w_i \in D_c$, $w_i \in W$, (2)

де g_v – кількість термів лінгвістичної змінної.

На рис. 1 множина D_c складається з документів, які знаходяться лівіше від вертикальної пунктирної лінії.

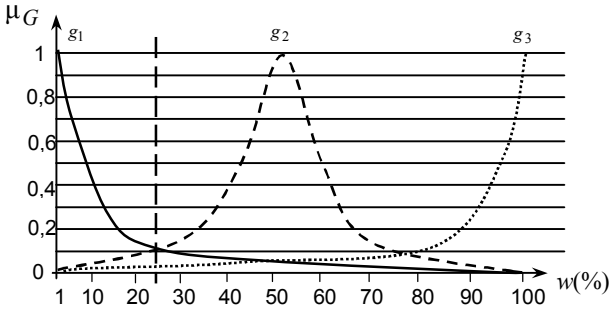

Рис. 1. Функції належності для лінгвістичної змінної “Множина для розширення запиту”

Масив D_c будується з пар виду:

$$\{ d_{l'}, \mu_{d_{l'}} \}, \ l' = \overline{1,l} \ ,$$

де l – кількість документів множини для розширення запиту D_c;

$d_{l'}$ – документ l' з множини D_c;

$\mu_{d_{l'}}$ – функція приналежності, яку набуває документ $d_{l'}$ з відповідного діапазону $w_i \in W$, $i = \overline{1,n}$.

Експертна оцінка проводиться один раз на тестовій колекції документів. У процесі роботи ІПС можливе коректування функції приналежності.

Так як документ $d_{s'} \in D$, $s' = \overline{1,s}$ і масив термів $T^{d_{s'}} = \{t_1, t_2, ..., t_f\}$ цього документа в інформаційно-пошукових системах – еквівалентні поняття, то ФП документа $\mu_{d_{s'}}$ і функція приналежності ПОД $\mu_{\text{ПОД}}$ даного документа також еквівалентні, та можна зробити висновок, що чим вище у документа ФП до множини документів для розширення запиту, тим більше вагомими є терми, які описують

даний документ, для вираження інформаційних потреб користувача.

В умовах автоматичної обробки текстів окремі показники змісту легко відокремити один від одного шляхом приписування їм ваг, відповідно до припущення про їхню важливість. Зважені показники забезпечують кращі результати пошуку, чим незважені [8].

Нехай C – колекція документів, T – множина усіх термів колекції, $T(d)$ – множина усіх термів документа $d \in C$, $tf(d,t)$ – число входжень терму t в документ d, $df(t)$ – число документів колекції C, що містять терм t. Визначимо вагу терму t в документі d через $w_t(d)$. Тоді $w_t(d)$ визначається як [8]:

$$w_t(d) = \frac{tf(d,t)\log(\frac{|C|}{df(t)})}{\sqrt{\sum_{t'} tf(d,t')\log(\frac{|C|}{df(t')})}}.$$

Якщо терм t не входить у документ d, тоді $w_t(d) = 0$.

Для модифікації запиту введемо поняття коефіцієнта важливості $\eta_{t_{f'}}(d_{l'})$ терму $t_{f'} \in T^{d_{l'}}$, що входить у документ $d_{l'} \in D_c$:

$$\eta_{t_{f'}}(d_{l'}) = \mu_{d_{l'}} w_{t_{f'}}(d_{l'}),\ d_{l'} \in D_c, t_{f'} \in T^{d_{l'}}, \quad (3)$$

де $w_{t_{f'}}(d_{l'})$ – вага $t_{f'}$-го терму документа $d_{l'}$;

$\mu_{d_{l'}}$ – функція приналежності документа $d_{l'}$ до масиву документів D_c.

Коефіцієнт важливості $\eta_{t_{f'}}(d_{l'})$ необхідний для визначення значимості терму $t_{f'}$ як кандидата на додаванні його в масив ключових слів $K(z_{m'})$, з якого відбиратимуться терми для розширення первинного запиту $z_{m'}$, $m' = \overline{1,m}$ користувача.

Підсумкове значення коефіцієнта важливості для однакових термів, що можуть ввійти в масив, обчислюється за формулою:

$$\eta_{t_{f'}\,рез} = \sum_{a=1}^{k} \eta_{t_{f'}}(d_a), \quad (4)$$

де $\eta_{t_{f'}}(d_a)$ – коефіцієнт важливості $\eta_{t_{f'}}(d_a)$ терму $t_{f'}$ документа $d_a \in D_c$;

k – кількість повторень окремо узятого слова в масиві ключових слів.

Масив $K(z_{m'})$ формується з усіх ключових термів кожного документа з множини для розширення запиту і сортується по зменшенню коефіцієнтів важливості.

Модифікований запит формується з вихідного запиту користувача $z_0 \in Z$ шляхом додавання термів $t_{f'} \in K(z_{m'})$ з найвищими коефіцієнтами важливості $\eta_{t_{f'}}(d_{l'}) \to \max$ і пропонується користувачу для подання в АІПС. Процес формування такого запиту є ітераційним. Користувач, ґрунтуючись на своєму досвіді, може як збільшити, так і зменшити кількість термів в результуючому запиті.

Вхідні та вихідні дані

Таким чином, у відповідності до визначеного математичного апарата, вхідними даними методики мають бути:

запит користувача $z_{m'} \in Z$, $m' = \overline{1,m}$ до ІПС;

множина документів $D = \{d_1, d_2, ..., d_s\}$, які проіндексовані ІПС;

масив релевантних документів $D_r \subset D$, які були відібрані на первинний запит та мають найбільш вагоме значення для потреб користувача, які виражені через його запит;

лінгвістична змінна "Множина для розширення запиту", множина термів $G = \{g_1, g_2, ..., g_v\}$ та базова множина $W = [w_1, w_n]$ лінгвістичної змінної.

Вихідними даними методики є модифікований запит, який було створено з урахуванням вихідного запита $z_{m'}$, $m' = \overline{1,m}$ користувача.

Алгоритм реалізації

Алгоритм реалізації методу складається з 2 етапів.

Етап №1. Визначення множини документів для розширення запиту. Цей етап складається з наступних кроків:

1.1. Визначення ФП нечітких термів лінгвістичної змінної;

1.2. Визначення множини документів $D_c \subset D_r$, терми з яких можуть виступати кандидатами до модифікації запиту, за формулою (2).

Етап №2. Побудова модифікованого запиту. На цьому етапі відбуваються наступні кроки:

2.1. Визначення коефіцієнта важливості $\eta_{t_{f'}}(d_{l'})$ для кожного ключового терму $t_{f'} \in T^{d_{l'}}$ документу $d_{l'} \in D_c$ за формулою (3);

2.2. Визначення $\eta_{t_{f'}\,рез}$ для однакових термів, за формулою (4);

2.3. Формування модифікованого запиту шляхом додавання термів $t_{f'} \in K(z_{m'})$ з найвищими коефіцієнтами важливості $\eta_{t_{f'}}(d_{l'}) \to \max$.

Висновки

Таким чином, запропоновано новий метод, який на відміну від існуючих не потребує побудови семантичної мережі чи тезауруса, що вигідно відрізняє його від інших підходів, про які говорилося вище. З причини того, що розширений запит будується з термів документів, релевантних інформаційним потребам користувача і добір найбільш релевантних документів для розширення запиту здійснюється на основі експертної інформації, це застерігає розширення запиту невірними словами, що в свою чергу виключає видачу документів, які не відносяться до інформаційних потреб користувача.

Наступним етапом дослідження, який дозволить використовувати даний метод в автоматизованих інформаційно-пошукових системах при вирішенні задач інформаційного пошуку, є оцінка ефективності запропонованого методу.

Література

1. Дубинский А.Г. Факторы, влияющие на качество информационного поиска // Системний аналіз та інформаційні технології: Зб. тез доп. Міжн. НПК студ., аспірантів та мол. вчених. Ч. 2. – К.: НТУУ "КПІ", 2001. – С. 43 – 48.

2. Гарант-Парк-Интернет: технологии анализа и поиска текстовой информации. – [Электр. ресурс]. – Режим доступа: http://www.research.metric.ru/cont.

3. Дубинский А.Г. Некоторые аспекты задачи построения автоматизированной поисковой системы // Научный сервис в сети Интернет: тез. докл. Всерос. НФ. – М.: МГУ, 1999. – С. 283 – 288.

4. Сизиков Е.В., Сошников Д.В. Онтологическая поисковая система Jewel для реализации интеллектуального поиска в Интернет- и Интранет-сетях // Электронный журнал "Труды МАИ". – 2002 . – № 7.

5. Герасимов Б.М., Субач І.Ю., Сергєєв О.Ю. Оцінка релевантності знайдених документів у розподілених інформаційно-пошукових системах // Автоматизація виробничих процесів. – 2005. – № 1. – С. 45 – 51.

6. Мелихов А.Н., Берштейн Л.С., Коровин С.Я. Ситуационные советующие системы с нечеткой логикой. – М.: Наука, 1990. – 272 с.

7. Тоценко В.Г. Методи та системи підтримки прийняття рішень. Алгоритмічний аспект. – К.: Наук. думка, 2002. – 381 с.

8. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. – М: Сов. радио, 1973. – 560 с.

Надійшла до редакції 20.02.2006

Рецензент: д-р техн. наук, проф. Б.М. Герасімов, Національний технічний університет України "КПІ", Київ.