navigation patterns // In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA, pages 31–36, 1999. 11. David Konopnicki. The w3ql query language and the w3qs system. Master's thesis, The Technion – Israel Institute of Technology, 1996. 12. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web // In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997. 13. Journey to the Internet's Unknown Regions, http://www.newsfactor.com/perl/story/17418.html. 14. http://www.sims. berkeley. edu/research/projects/how-much-info/internet.html.

УДК 681

О.Я. Тарас
Національний університет "Львівська політехніка",
кафедра "Інформаційні системи та мережі"

ОГЛЯД ІСНУЮЧИХ АЛГОРИТМІВ ТА МОДЕЛЕЙ ПОШУКУ У WEB

© *Тарас О.Я., 2006*

The review of algoritmes and searching modes are represented in this artickle.

Подано огляд алгоритмів та моделей пошуку, що застосовуються в пошукових системах.

ВСТУП

Пошукові і не тільки пошукові системи Інтернет настільки популярні сьогодні, що люди проводять години, обговорюючи переваги і недоліки, алгоритми та програми для пошуку повнотекстової інформації на тих або інших носіях. І весь цей час не вщухають гарячі суперечки між фахівцями та користувачами. Перші – прихильники суто "механічних" машин, пошукових систем, що обчислюють строгі логічні запити і підтримують усікання слова праворуч "зірочкою"; вони переконані, що краще всяких алгоритмів сформулюють, що ж їм потрібно знайти. Інші – навпаки, намагаються віддати алгоритмам розвідувача всі магічні перетворення вихідного запиту і не замислюватися про те, що ж там відбувається усередині. Обидві точки зору мають право на існування. Ми ж розглядатимемо цю проблему з позиції перших.

Бурхливий ріст обсягу інформації в Інтернет робить пошук незамінним методом доступу до цієї інформації. Можна виділити дві основні форми пошуку в Інтернет:

- *Використання пошукових систем*, що збирають відомості про ресурси, доступні в Інтернет, і організовують пошук за цією інформацією, як за повнотекстовою базою даних. Прикладами таких систем є – Altavista, Google, Яndex тощо.
- *Використання Інтернет-каталогів*, у яких інформація про обрані ресурси Інтернет класифікована за тематичними ознаками. Такі каталоги існують не тільки в електронному виді (List.Ru або Yahoo!), але також видаються і у вигляді друкованих видань – таких як, наприклад, "Жовті сторінки Інтернет''.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Природа Інтернет обумовлює ряд важливих факторів, які необхідно враховувати при розгляді задач пошуку:

- *Величезний обсяг доступної інформації*

Так, станом на лютий 2000 року в Інтернет було опубліковано більш ніж мільярд сторінок, і це число збільшується експоненційно [1].

- *Високий відсоток тимчасової інформації*

У зв'язку з високою динамікою розвитку Інтернет, інформаційні ресурси дуже часто з'являються, щезають, змінюють свої місцерозташування або зміст. Відповідно до деяких оцінок щомісяця змінюється близько 42 % інформації [4].

- *Неконтрольована якість інформації*

Відсутність контролю спричиняє до появи некоректної (наприклад, уже застарілої), помилкової або неповної інформації: найчастіше це викликають граматичні помилки, помилки оцифровки тощо.

- *Різнорідність інформації*

Крім різних форматів представлення інформації, до цієї групи особливостей відноситься також і те, що для представлення інформації використовується безліч різних мов і навіть алфавітів.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ

Наведено огляд існуючих алгоритмів та моделей пошуку з метою виявлення кола найактуальніших задач для подальшого дослідження. Стаття має оглядово-аналітичний характер і є першим кроком для подальших практичних та наукових розробок.

Пошукові системи

Величезний обсяг доступної в Інтернет інформації робить пошукові системи незамінним інструментом [4]. Більшість з них належить до одного з двох класів:

- *Багатоцільові системи*:

Такі системи (Altavista, Google, Infoseek, Яndex і т. п.) призначені для пошуку інформації з будь-яких запитів. Для виконання цієї задачі вони намагаються проіндексувати усю доступну в Інтернет інформацію.

- *Спеціалізовані системи*:

На відміну від багатоцільових систем, такі системи призначені для відповідей на запити, що належать до деякої спеціалізованої області. Наприклад, інформацію про життя міста Львова і основні події можна шукати на www.misto.ridne.net, прогноз погоди – на www.weather.com. Спеціалізована пошукова система здійснює пошук за значно меншою кількістю ресурсів, ніж будь-яка популярна багатоцільова пошукова система. Однак, цей факт має ряд позитивних наслідків для спеціалізованих систем [5]:

Інформація, що не належить до спеціалізації даної пошукової системи, не потрапляє в її індекс.

Можливе застосування більш обчислювально трудомістких методів пошуку.

Можливе залучення експертів у відповідній області, а також підтримка сервісу рекомендацій ресурсів користувачами системи. А як наслідок, підвищення якості і повноти колекції.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Тому найчастіше пошук у відповідному запитові спеціалізованій пошуковій системі швидше і краще задовольняє інформаційні потреби користувача.

Водночас, через спеціалізацію таких пошукових систем вибір конкретної системи для виконання пошуку є досить складною задачею. Для вирішення цієї проблеми різні пошукові системи чи інтелектуальні агенти пропонують свої власні підходи та методи, що базуються на розроблених ними або по-новому реалізованих алгоритмах. Одним з них є побудова тематичних карт або мереж заданої колекції документів. Цей підхід пропонує група розробників [1]. Такий підхід дуже трудомісткий і не завжди працює через обмеженість побудованих вручну описів. Автоматична побудова таких описів є предметом сучасних досліджень [5].

Зазначимо, що у межах цієї роботи ми не розглядаємо системи і методи пошуку, що враховують інформацію про структуру даних, такі як методи роботи зі слабоструктурированною інформацією [2].

Індексно-пошукові системи

Найважливішою відмінністю пошукових систем для пошуку в Інтернет від класичних систем інформаційного пошуку є необхідність обслуговування всіх запитів без реального доступу до ресурсів на момент виконання запиту. Інакше необхідно або зберігати свіжу локальну копію всіх ресурсів (що занадто дорого), або відвідувати ресурси під час виконання запиту (що сповільнює пошук).

Тому у системах пошуку в Інтернет усі запити обслуговуються на основі вмісту індексу, що містить деякі описи відомих даній пошуковій системі ресурсів. Для збору відомостей про доступні ресурси, що потім використовуються для побудови індексу, зазвичай використовуються так звані мережні роботи – програми, що починаючи з деякої Інтернет-сторінки, рекурсивно обходять ресурси Інтернет, витягаючи посилання на нові ресурси з одержуваних документів [2, 3, 5].

Проте, навіть наймогутніші пошукові системи не можуть проіндексувати усю доступну в Інтернет інформацію.

Розподілені пошукові системи

З метою підвищення продуктивності і надійності більшість сучасних багатоцільових систем мають уже не централізовану, а розподілену архітектуру [1].

В останні роки активно досліджувалася можливість застосування розподілених архітектур до пошукових систем [2, 4, 5]. У розподілених пошукових системах єдиний індекс розбивається на кілька окремих частин (колекцій) за деяким принципом, причому усередині різних колекцій можуть використовуватися різні методи виконання пошуку. При створенні розподіленої пошукової системи необхідно вирішити ряд питань [3]:

- *Як маршрутизуваті запити?*

Для зниження навантаження на мережу і підвищення ефективності операція пошуку виконується не у всіх колекціях, а тільки в деякій їхній частині. Цей процес називається маршрутизацією запитів (query routing) [4, 5].

- *Як поєднувати відповіді?*

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Процес об’єднання відповідей від окремих частин індексу в єдину відповідь системи називається злиттям результатів (database fusion) [3, 4].

- *Які протоколи використовувати?*

Вибір протоколу обміну даними у межах системи впливає не тільки на потенційну функціональність, але і на відкритість системи [1]. Прикладами існуючих протоколів є STARTS , Z39.50.

ФОРМУВАННЯ ЦІЛЕЙ

Задачі інформаційного пошуку

Центральна проблема інформаційного пошуку формулюється просто – допомогти користувачеві знайти ту інформацію, у якій він зацікавлений [5]. На жаль, описати інформаційні потреби користувача зовсім не так просто. Переважно цей опис формулюється як деякий запит, що є деяким набором ключових слів, що характеризує потреби користувача.

Класична задача інформаційного пошуку, з якої й почався розвиток цієї галузі, – це пошук документів, що задовольняють запит, у межах деякої статичної (на момент виконання пошуку) колекції документів. Наприклад, ця задача вирішується у межах більшості сучасних довідкових систем, таких як довідкова система з операційної системи Windows.

Однак за тридцять років досліджень список задач інформаційного пошуку значно розширився і тепер містить питання моделювання, класифікації і кластеризації документів, проектування архітектур пошукових систем і користувацьких інтерфейсів, мови запитів тощо [1].

Крім класичної задачі інформаційного пошуку, в цій роботі ми також торкаємося таких задач:

- *Кластеризація документів*

Метою кластеризації документів є автоматичне виявлення груп семантично схожих документів серед заданої фіксованої безлічі документів [6]. Зазначимо, що групи формуються тільки на основі попарної схожості описів документів, і жодні характеристики цих груп не задаються заздалегідь.

- *Класифікація документів*

На відміну від задачі кластеризації, метою цієї задачі є визначення для кожного документа однієї або декількох із заздалегідь заданих категорій, до яких цей документ належить. Особливістю задачі класифікації є припущення, що безліч класифікованих документів не містить “сміття’’, тобто кожний з документів відповідає якій-небудь із заданих категорій.

Частковим випадком задачі класифікації є задача тематичної класифікації. Тут кожна категорія – це деяка тематика, а ціль класифікації – визначити тематику документа [2, 4].

- *Фільтрація документів*

Як і в задачі класифікації, метою задачі фільтрації є розбивка безлічі документів на категорії. Однак цих категорій тільки дві – ті документи, що задовольняють заданий критерій, і ті, що не задовольняють.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Однією з найбільш важливих є задача тематичної фільтрації документів, тобто автоматичного визначення документів, що відповідають заданій тематиці, за рахунок відсіювання інших документів [5, 3].

Незважаючи на деяку схожість формулювань цих задач, вони сильно відрізняються. Як наслідок, методи, успішно застосовувані для розв'язання однієї з цих задач, найчастіше показують не кращі результати при використанні їх для розв'язання іншої задачі.

ОСНОВНИЙ МАТЕРІАЛ

Основну увагу звернемо на класичні моделі пошуку, оскільки більшість пошукових машин та спеціалізованих агентів дотримуються саме цих моделей.

Класичні моделі інформаційного пошуку розглядають документи як набори ключових слів, що зустрічаються у цих документах і які надалі називатимемо термами. Терм – це просте слово, семантика якого допомагає описати основний зміст документа.

Формально опис будь-якої моделі інформаційного пошуку складається з чотирьох частин [16]:

D – безліч використовуваних типів представлень документів;

Q – безліч використовуваних типів описів інформаційних потреб користувача, тобто запитів;

F – загальний каркас, у межах якого відбувається моделювання описів документів і запитів, а також опис взаємозв'язків між ними

$R(q, d_i)$ – функція ранжування, що парі документ/запит зіставляє деяке дійсне число

Моделі інформаційного пошуку поділяються на три класи:

- *Теоретико-множинні моделі*

Моделі цього класу використовують як каркас теорію множин. Класичний приклад – булева модель. У межах цієї моделі документи і запити представляються у вигляді безлічі термів.

- *Імовірнісна модель*

Каркасом для таких моделей є теорія імовірностей. Як оцінку релевантності документа запитові користувача використовується імовірність того, що користувач визнає документ істинно релевантним.

- *Алгебраїчна модель*

У межах алгебраїчних моделей документи і запити описуються у вигляді векторів у багатомірному просторі. Каркасом для таких моделей є алгебраїчні методи.

У межах кожного з класів було запропоновано безліч альтернативних моделей. Незважаючи на ряд недоліків, на практиці класичні теоретико-множинні моделі досить популярні в силу своєї простоти. Хоча імовірнісні моделі пропонують найбільш природний спосіб формально описати проблему інформаційного пошуку, їхня популярність відносно невелика. Найбільш популярними є алгебраїчні моделі, оскільки їхня практична ефективність зазвичай виявляється вищою. Загалом пропоновані останнім часом нові моделі інформаційного пошуку найчастіше є гібридними і мають властивості моделей різних класів.

У межах цієї роботи ми спираємося на алгебраїчні моделі інформаційного пошуку, що описані нижче.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Векторна модель

Векторна модель є класичним представником класу алгебраїчних моделей. У межах цієї моделі кожному терму t_i в документі d_i (і запиті q) зіставляється деяка ненегативна вага $w_{i,j}$ (w_i для запиту). Отже, кожен документ і запит може бути представлений у вигляді k-мірного вектора

$$\overrightarrow{d_j} \xrightarrow{def} \left(w_{1j}, w_{2j}, ..., w_{kj}\right)$$

де k – загальна кількість різних термів у всіх документах.

Відповідно до векторної моделі, близькість документа d_i до запиту q оцінюється як кореляція між векторами їхніх описів. Ця кореляція може бути обчислена, наприклад, як скалярний добуток відповідних векторів описів.

Ваги термів можна обчислювати безліччю різних способів [70].Один з можливих підходів – використовувати як вагу терма $w_{i,j}$ в документі d_i нормалізовану частоту його використання $freq_{ij}$ у межах даного документа, тобто:

$$w_{ij} \xrightarrow{def} tf_{ij} = \frac{f\ \mathrm{req_{ij}}}{\max_i f\ \mathrm{req_{ij}}}$$

Однак цей підхід не враховує, наскільки часто даний терм використовується в інших документах колекції, тобто дискримінаційну силу терма. Тому у випадку, коли доступна статистика використань термів у колекції, краще працює інша схема обчислення ваг 1.6:

$$w_{ij} \xrightarrow{def} tfid f_{ij} = t f_{ij} * \log\frac{n_j}{N}$$

де n_i позначає число документів, у яких використовується терм t_j, а N – загальне число документів у колекції.

Латентно-семантичний аналіз

Латентно-семантичний аналіз (LSA) – це теорія і метод для витягу контекстно-залежних значень слів за допомогою статистичної обробки великих наборів текстових даних [5]. Протягом декількох останніх років цей метод не раз використовувався як в області пошуку інформації [4, 5], так і в задачах фільтрації і класифікації [3].

Латентно-семантичний аналіз ґрунтується на ідеї, що сукупність усіх контекстів, у яких зустрічається і не зустрічається дане слово, задає безліч обопільних обмежень, що значною мірою дозволяють визначити подібність значеннєвих значень слів і безлічі слів між собою.

Як вихідну інформацію LSA використовує матрицю терми-на-документи, що описує використовуваний для навчання системи набір даних. Елементи цієї матриці містять частоти використання кожного терма в кожному документі.

Найбільш розповсюджений варіант LSA заснований на використанні розкладання матриці за сингулярним значенням (SVD)[2, 5].Використовуючи SVD, величезна вихідна матриця розкладається в безліч з k, зазвичай від 70 до 200, ортогональних матриць, лінійна комбінація яких є непоганим наближенням вихідної матриці.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Більш формально, відповідно до теореми про сингулярне розкладання [2, 5], будь-яка дійсна прямокутна матриця X може бути розкладена в добуток трьох матриць:

$$X = U\Sigma V^T$$

таких, що матриці U і V – ортогональні, а Σ – діагональна матриця, значення на діагоналях якої називаються сингулярними значеннями матриці X.

Таке розкладання має цікаву особливість: якщо в Σ залишити тільки k найбільших сингулярних значень, а в матрицях U і V тільки відповідні до цих значень стовпці, то добуток матриць, що Σ_{Lsa} вийшли U_{Lsa}, V_{Lsa}, і буде найкращим наближенням вихідної X матриці матрицею k рангу :

$$X \approx \hat{X} = U_{Lsa}\Sigma_{Lsa}V_{Lsa}$$

Основна ідея латентно-семантичного аналізу полягає в тому, що якщо як X використовувалася матриця терми-на-документи, то матриця $\hat{X}$, що містить тільки k перших лінійно незалежних компонент X, відбиває основну структуру асоціативних залежностей, що є присутні у вихідній матриці і водночас не містить шуму.

Отже, кожен терм і документ представляються за допомогою векторів у загальному просторі розмірності k (так званому просторі гіпотез). Близькість між будь-якою комбінацією термів і/або документів може бути легко обчислена за допомогою скалярного добутку векторів.

Вибір найкращої розмірності k для LSA – відкрита дослідницька проблема. В ідеалі k має бути досить велике для відображення всієї реально існуючої структури даних, але водночас досить малим, щоб не захопити випадкові і маловажливі залежності. Якщо обране k занадто велике, то метод втрачає свою цінність і наближається за характеристиками до стандартних векторних методів. Занадто маленьке k не дозволяє уловлювати розходження між схожими словами або документами. Дослідження показують, що з ростом k якість спочатку зростає, а потім починає падати [3].

Альтернативу формальним лінгвістичним підходам становить клас статистичних методів аналізу тексту, і саме вони використовуються в комерційних системах для вирішення таких задач, як автоматичне реферування, тематична класифікація і кластеризация текстів, змістовний пошук, які можна розглядати в комплексі як задачу тематичного аналізу. Статистична інформація про окремі лексичні одиниці легко видобувається з тексту і є всі підстави думати, що вона адекватно відображає його зміст загалом. Непряме підтвердження цьому можна знайти в нейропсихологічних дослідженнях, які встановили, що аналіз друкованого тексту, спираючись на зорове просторове (а не на лінійне слухове) сприйняття, реалізується переважно правою півкулею мозку, що використовує асоціативну статистичну модель [2, 3]. Логічний “лівопівкулевий” аналіз, моделюванням якого по суті займається формальна лінгвістика, необхідний лише в окремих "важких" місцях тексту, що несуть нову інформацію і вимагають детального осмислення.

ПРАКТИЧНА ЧАСТИНА

Основною задачею статистичного аналізу є дослідження розподілу лексики у межах різних одиниць тексту – зразків і фрагментів, що бажано проводити з залученням більш загальної статистичної моделі. Описуваний підхід розв’язання задачі представляє конструк-

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

тивний розвиток ідей, викладених у роботі [4], і претендує на спробу моделювання способів обробки інформації правою півкулею людського мозку. Ключовим моментом підходу є можливість використовувати на визначеному етапі аналізу асоціативну семантично-статистичну модель, сформовану на базі самого досліджуваного тексту.

АНАЛІЗ НАДФРАЗОВОЇ СТРУКТУРИ ТЕКСТУ НА ОСНОВІ АСОЦІАТИВНОЇ СЕМАНТИЧНОЇ МЕРЕЖІ

В основі підходу лежить інтегральне представлення змісту тексту у формі асоціативної семантичної мережі, описане в роботі [4]. Семантична мережа – це безліч понять тексту – слів і словосполучень, асоціативно зв'язаних між собою. Як критерій зв'язности пропонується використовувати частоту спільної зустрічності понять у зразках тексту. Швидкий алгоритм виділення зв'язних словосполучень, що представляють цілісні поняття мережі, ґрунтується на аналізі частоти зустрічності ланцюжків слів різної довжини і їхнього входження один в один, що може бути зроблений із застосуванням багаторівневої нейроподібної структури [4]. Найважливішою властивістю семантичної мережі є диференціація зв'язків за вагами, що відбивають ступінь змістовної зв'язності понять. Зв'язок від поняття i до поняття j пропонується характеризувати вагою w_{ij}, що у найпростішому випадку визначається як

$$w_{ij} = \frac{f_{ij}}{f_j}, \tag{1}$$

де f_{ij} – частота спільної зустрічності понять у зразках тексту, а f_i – власна частота зустрічності поняття в тексті. Як видно, вага зв'язку відбиває умовну імовірність того, що при згадуванні в тексті поняття i мова також йде про поняття j. У загальному випадку вага зв'язку між парою понять може враховувати їхній зв'язок через треті поняття, що можна представити як спрощену модель механізму реорганізації інформації у сні [4].

Розглянемо застосування такої семантичної мережі для тематичного аналізу тексту, заснованого на виділенні цілісних фрагментів, зв'язаних загальним змістом – надфазних єдностей (НФЕ) [5]. Можна вважати, що кожна НФЕ характеризується головною темою, а кожній з тем відповідає ряд НФЕ в тексті. Окремі НФЕ для різних тем можуть перетинатися або включатися одна в одну, що відбиває ієрархічну тематичну структуру тексту – його комунікативна побудова в процесі породження автором. Через те, що комунікативне членування тексту при сприйнятті спирається на сформовану семантичну модель реципієнта [6], розв'язання задачі надфразового аналізу в принципі не однозначне і визначається структурою використовуваної моделі. Як така в найпростішому випадку може бути використана семантична мережа самого досліджуваного тексту, а в більш загальному – мережа, попередньо створена на базі еталонних текстів. Вважаючи, що кожна тема відповідає одному з понять семантичної мережі, задачу виділення НФЕ можна сформулювати як задачу пошуку фрагментів тексту, близьких "за змістом" до відповідних понять. При цьому як опис теми можна використовувати набір зв'язків поняття в мережі і прийняти, що при згадуванні поняття в тексті мова також йде про усі зв'язані з ним поняття, у ступені,

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

пропорційному вагам відповідних зв'язків. Щоб оцінити приналежність окремих зразків до тем, введемо поняття рівня активації i-го елемента на зразок з номером t:

$$w_i^*(t) = \sum_j \sigma_j(t) w_{ij} \Big/ \sum_j \sigma_j(t) \text{ , j=1..I,} \quad (2)$$

де $\sum_j \sigma_j(t)$ – кількість слів у зразку t; I – кількість елементів мережі; $\sigma_j(t) = \{1$, якщо поняття присутнє в зразку ; 0 – у протилежному випадку $\}$.

Таке визначення означає, що поняття, що зустрілося в зразку, підвищує рівень активації кожного з інших понять мережі на величину, пропорційну відповідній вазі зв'язку. У результаті можуть бути значно активізовані поняття, що мають сильні зв'язки з поняттями зі зразка. Ця властивість забезпечує стійкість тематичного аналізу до використовуваної лексики за рахунок того, що при аналізі локальних ділянок тексту використовується сукупна статистична інформація про зміст текстів, що формували семантичну мережу, що апріорі має більш високу вірогідність.

Тематичну приналежність ділянок тексту характеризує сукупний рівень активації елементів на інтервалі $\Delta T = (t, t+m-1)$:

$$w_i^*(t,m) = \sum_k w_i^*(t+k), k = 0..m. \quad (3)$$

Як видно, рівень активації показує ступінь "насиченості" фрагмента тексту інформацією, що належить до теми, представленої і-им елементом. Можна вважати, що динаміка рівнів активації понять на тимчасовій осі тексту загалом відбиває його комунікативну структуру з погляду сприймаючого, що використовує як модель предметної області семантичну мережу. У цьому випадку з'являється можливість виділення НФЕ для кожної з тем як ділянок з високим рівнем активації. Як НФЕ вибирається послідовність зразків з найбільш тривалого інтервалу ΔT, що забезпечує виконання таких умов :

$$w_i^*(t) \geq \varpi,\ w_i^*(t+m-1) \geq \varpi_i \quad (4)$$

$$/\exists\ \Delta T' \subset \Delta T,\ |\Delta T'| > \tau\ \ w_i^*(t) < \omega \text{ для } \forall\ t' \in \Delta T'.$$

Тут ω – параметр, що задає граничне значення рівня активації, перевищення якого дозволяє говорити про віднесеність інформації в зразку до i-ї теми. Параметр τ накладає обмеження на зв'язність НФЕ. Його значення вказує максимальну кількість зразків, протягом яких допускається відхилення в змісті тексту від теми, що відповідає елементові. Доцільно ввести адаптацію порога ω до рівня активації елементів виду

$$\omega = \omega(t) = \sum_j \sigma_j(t)\, w^*_j(t) / \sum_j \sigma_j(t),\ j = 1..I. \quad (5)$$

Отже, значення порога приймається рівним середньому рівневі активації на поняттях, що входять у пропозицію, що аналогічно введенню літерального гальмування між елементами мережі. Можлива також додаткова адаптація параметра τ до сукупного рівня активації поняття в інтервалі НФЕ $\Delta T = (t_0, t_0 + m-1)$, наприклад:

$$\tau = \tau(t, i) = \ln w^*_i(t, t-t_0). \quad (6)$$

Це відбиває той факт, що при збільшенні тривалості НФЕ обмеження на зв'язність може трохи слабшати, а для короткого НФЕ розумно допускати лише короткотермінове відхилення від теми. Загалом, варіювання τ дозволяє регулювати детальність надфразового аналізу.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

РЕАЛІЗАЦІЯ ПРИКЛАДНИХ ФУНКЦІЙ ТЕМАТИЧНОГО АНАЛІЗУ ТЕКСТУ

Результатом надфразового аналізу є виділення безлічі непересічних НФЕ для кожної з тем тексту : $\{\Delta T_{i,1}, \Delta T_{i,2}, \ldots \Delta T_{i,P(i)}\}$, $\Delta T_{i,p} = (t_{i,p}, t_{i,p} + m_{i,p} -1)$, де P(i) – кількість НФЕ за і-ою темою . При цьому як оцінку інформативності НФЕ для теми може використовуватися сукупний рівень активації елемента на НФЕ-інтервалі – реферативна вага:

$$r_{i,p} = w^*_i(t_{i,p}, m_{i,p}). \quad (7)$$

Тепер можливо оцінити “значимість” теми для тексту загалом як сукупну реферативну вагу за всіма НФЕ, що належать до:

$$W^*_i = \sum_p r_{j,p} , p = 1.. P(i). \quad (8)$$

Як видно, подібний спосіб оцінки враховує тільки “невипадкові” згадування в тексті відповідного поняття і зв’язаних з ним.

Ранжирування тем на основі сукупних реферативних ваг дозволяє охарактеризувати ступінь віднесеності тексту до кожної з тем і виділити головні. Послідовність НФЕ з високою реферативною вагою, що належать до однієї теми і розташованих у порядку проходження в тексті, може інтерпретуватися як тематичний реферат тексту. Сформувати загальний реферат тексту можна з найбільш вагомих НФЕ за найбільш значимими темами. Зважаючи на те, що метою реферування є вміщення найбільшої кількості інформації в обмежений обсяг, доцільною є оцінка значимості кожного НФЕ з обліком перетину з НФЕ інших тем. Так, наприклад, якщо одне НФЕ містить у собі інше, те його реферативна вага повинна бути збільшена з урахуванням ваги другого НФЕ і його теми. Тоді загальну реферативну вагу НФЕ за всіма темами можна визначити:

$$R_{i,p} = \sum_j \sum_q r_{j,q} W^*_j |\Delta T_{i,p} \cap_\Delta T_{j,q}| / |\Delta T_{j,q}|, \; q = 1..Q(j), \; j = 1..I, \; j \neq i, \quad (9)$$

де $|\Delta T_{i,p} \cap_\Delta T_{j,q}|$ – довжина інтервалу перетинання ділянок НФЕ *i* -го і *j* -го понять.

Остаточно в реферат вибираються НФЕ з максимальною загальною реферативною вагою, за порядком проходження в тексті. При цьому пересічні НФЕ поєднуються в одне.

Можливі й інші стратегії компонування реферату.

Як видно, описаний підхід можна застосувати і до аналізу окремого тексту без наявності апріорної інформації, і до аналізу тексту на основі попередньо сформованої семантичної мережі. У другому випадку відбувається фільтрація інформації тексту, що належить до тем з еталонної мережі, що представляє одну з форм реалізації функції змістовного пошуку. Даний принцип є реалізованим в окремих інтелектуальних агентах, що застосовуються при пошуку в Web.

ВИСНОВОК

Описані алгоритми досліджені і реалізовані в компанії “Гарант-Парк-Інтернет” у ході розробки програмної бібліотеки *GPTopMining*. Універсальність підходу й однорідність використаних способів обробки інформації дозволили продемонструвати високу швидкість аналізу в сполученні з якістю, що перевершує існуючі на ринку комерційні розв’язання задач тематичного аналізу текстів. Отримані результати є базою для подальшого аналізу з метою оптимізації пошуку інформації у Web.

1. Мельчук И.А Опыт теории лингвистических моделей “Смысл-Текст”. Семантика, синтаксис. – М.: Школа “Языки русской культуры”, 1999. 2. Глезерман Т.Б. Психофизиоло-

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

гические основы нарушений мышления при афазии. – М.: Наука, 1996. – 230 с. 3. Брагина Н.Н., Доброхотова Т.А. Функциональные асимметрии человека. – М: Медицина, 1981. – 287 с. 4. Харламов А.А., Ермаков А.Е., Кузнецов Д.М. Технология обработки текстовой информации с опорой на семантическое представление на основе иерархических структур из динамических нейронных сетей, управляемых механизмом внимания // Информационные технологии. – 1998. – № 2. – С. 26–32. 5. Орлова Л.В. Структура сверхфразового единства в научных текстах. – К.: Наукова думка, 1998. – 154 с. 6. Ахутина Т.В. Порождение речи. Нейролингвистический анализ синтаксиса. – М.: МГУ, 1989. – 215 с.

УДК 681.3

Д.О. Тарасов
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ФОРМАЛЬНІ МОДЕЛІ СИСТЕМ ЗАХИСТУ ІНФОРМАЦІЇ РЕЛЯЦІЙНИХ БАЗ ДАНИХ

© *Тарасов Д.О., 2003*

This paper describes the formal models of relational databases information protection system. We propose new formal model of relational databases information protection system.

Розглянуто формальні моделі систем захисту інформації реляційних баз даних та методи їх реалізації. Запропоновано нову формальну модель захисту інформації реляційної БД.

1. ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ

Використання технологій баз даних дозволяє ефективно та швидко аналізувати інформацію, формалізувати процес проектування інформаційних систем (ІС), швидко опрацьовувати великі об’єми інформації. Для найбільш розповсюджених баз даних – реляційних баз даних (БД) створено стандартизований інструментарій систем управління базами даних (СУБД) [1].

Питанням захисту інформації БД, засобам захисту інформації СУБД постійно приділяється увага [1, 3, 4, 6–8, 10]. Це зумовлено інформацією про знайдені недоліки захисту СУБД, появою нових методів аналізу інформації, винекненням нових задач захисту інформації. Для вирішення проблеми захисту інформації у БД необхідний комплексний підхід, який базується на формальних моделях системи захисту інформації (СЗІ) [12–14].

2. АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

У дослідженнях СЗІ БД розглядають такі аспекти захисту:

- механізми транзакцій;
- забезпечення цілісності реляційних БД;
- принципи побудови та використання БД з примусовим контролем доступу (mandatory access control, MAC);

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua