НАНУ, 2010. – Т. 1, № 8. – С. 120–124. 14. Слуцкер А. И. Характеристики элементарных актов в кинетике разрушения металлов / А. И. Слуцкер // Физика твердого тела. – 2004. – Т. 9, №46. – С. 1606–1613. 15. Стащук М. Розрахунок зміщення електродного потенціалу, зініційованого пружним полем, на межі еліптичного отвору із середовищем / М. Стащук, Л. Журавчак, М. Дорош // Проблеми корозії та протикорозійного захисту матеріалів: у 2-х т. / Спецвип. журн. «Фізико-хімічна механіка матеріалів». – Львів: ФМІ ім. Г. В. Карпенка НАНУ, 2010. – Т. 1, №8. – С. 49–54. 16. Юзевич В. Моделювання корозійних процесів у системі «метал–електроліт» з урахуванням дифузійного імпедансу / В. Юзевич, І. Огірко, Р. Джала // Фізико-математичне моделювання та інформаційні технології. – 2011. – Вип. 13. – С. 173–181. 17. Юзевич В. М. Діагностика матеріалів і середовищ. Енергетичні характеристики поверхневих шарів /В. М. Юзевич, П. М. Сопрунюк. – Львів: ФМІ ім. Г. В. Карпенка НАНУ; Сполом, 2005.–292 с.

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ОЦЕНКИ СОСТОЯНИЯ ОБЪЕКТОВ ИЗ СТАЛИ В МОРСКОЙ СРЕДЕ ИЗ СЕРОВОДОРОДОМ С ИСПОЛЬЗОВАНИЕМ АЛГЕБРЫ АЛГОРИТМОВ

Разработана информационная технология отбора и обработки данных относительно оценивания энергетических характеристик междуфазных слоев и активационных процессов, которые характеризуют металл (сталь) и динамику коррозионных процессов вблизи вершины каверны в морской воде из сероводородом.

INFORMATION TECHNOLOGY ASSESSMENT OBJECT STATE OF STEEL IN THE MARINE ENVIRONMENT OF HYDROGEN SULFIDE USING ALGEBRA ALGORITHMS

Information technology of selection and working of data in relation to the evaluation of power descriptions of interface layers and activating processes that characterize a metal (steel) and dynamics of corrosive processes near-by the top of cavity in marine water with the sulphuretted hydrogen is presented.

Стаття надійшла 12.10.2012

УДК 519.76: 81.37

О.С. Ляшко, І.З. Миклушка
Українська академія друкарства

ПОДАННЯ СЕМАНТИКИ ТЕКСТУ ЧЕРЕЗ ЙОГО СТАТИСТИЧНІ ПОКАЗНИКИ

Проаналізовано методи поданих текстових документів в інформаційно-пошукових системах. Окреслено можливості відтворення змісту тексту через його статистичні показники.

Семантика тексту, статистичні показники, наукові тексти, семантичний пошук, векторна модель тексту

В умовах необхідності здійснення пошуку в постійно зростаючому масиві текстової інформації наукового характеру особливо гостро постає питання попереднього опрацювання текстових масивів. Зокрема, актуальною є проблема автоматичного аналізу повнотекстових документів, автоматичної класифікації, виявлення тематики, автоматичного реферування документів та установлення їх семантики [7].

Слід зазначити, що сьогодні жодна із систем аналізу текстових даних не забезпечує всіх поставлених до неї вимог стосовно повноти, точності та рівня автоматизованості. Дана ситуація пов'язана з тим, що наразі немає достатньо адекватних моделей таких систем [3]. Це у сфері опрацювання текстових даних зумовлює постійне виникнення нових більш досконалих моделей подання текстів для використання їх у подальшому в інформаційно-пошукових системах (ІПС).

Основними проблемами, які ставляться перед ІПС при опрацюванні текстових документів, є: зменшення обсягу тексту при збереженні його семантики; автоматичне реферування [7]; категоризація; побудова семантичних зв'язків; створення статистичного портрета.

Нині серед традиційних методів аналізу [4] можна виділити статистичні та семантичні. З огляду на доволі велику складність алгоритмів проведення семантичного аналізу, поки що ці методи не здобули значного розповсюдження і використовуються переважно для дослідницьких цілей. На практиці ж статистичні методи достатньо популярні в багатьох ІПС завдяки можливості опрацьовувати великі масиви даних і вирішувати різноманітні завдання щодо опрацювання отриманих за результатами аналізу даних [3]. Завданням статистичної обробки тексту є математичний опис його мовних фактів та явищ, зв'язків між ними, отримання набору моделей для вирішення визначених лінгвістичних задач [5].

Метою цієї статті є розгляд питання виділення семантики тексту з результатів його статистичного аналізу, тобто перетворення математичного опису тексту в його смислову сутність. Це, у свою чергу, дозволить, зменшивши затрати на побудову семантичної моделі тексту, удосконалити можливості ІПС щодо видачі релевантних результатів пошуку по ньому. Основним завданням при цьому має бути коректний вибір ознак, за якими здійснюватиметься формування статистичного портрета тексту.

Векторна модель подання документів в ІПС

Інформаційний пошук – це процес пошуку у великій колекції деякого неструктурованого матеріалу (як правило, документа), що задовольнятиме інформаційну потребу [11]. Насправді неструктурованих документів наукового типу практично не зустрічаємо, адже існують певні характеристики, які дозволяють визначити текст як науковий. Однією з таких характеристик є чітка структурна побудова поданої в документі інформації. З огляду на це можна зазначити, що інформаційний пошук у наукових текстах являє собою процес отримання з масиву текстових документів релевантних даних як структурованого, так і неструктурованого типу.

Найпростішим способом опрацювання текстових документів ІПС статистичними методами є просте індексування, у результаті якого формується матриця двійкових значень наявності того чи іншого терміна в будь-якому документі. У реальному масиві документів створення такого індексу не є складним заняттям, проте об'єм пам'яті, який займатиме даний індекс, буде досить наближеним до об'єму пам'яті, що займає сам масив (а створення такого масиву, відповідно, втрачає будь-який зміст). При цьому більшість елементів даного масиву займатиме нульові значення. Більш оптимальним з точки зору використання пам'яті є інвертований індекс, що показує список документів, в яких зустрічається той чи інший термін.

У результаті кожного пошукового запиту ІПС видає документ з двійковою ймовірністю релевантності (релевантний або ні). Оцінкою дії кожної ІПС є такі показники, як точність (P) і повнота (R), які підраховуються за формулами

$$P = tp/(tp + fp);\ R = tp/(tp + fn),$$

де *tp* – релевантні документи, відтворені в результатах пошуку; *fp* – нерелевантні документи, отримані внаслідок пошуку; *fn* – релевантні документи, не відтворені в результатах пошуку.

Ураховуючи те, що дана модель (індексування) не відображає частоти вживання того чи іншого терміна в даному тексті (абсолютну, відносну частоти), а тільки наявність або відсутність терміна, результати, які відповідають пошуковому запиту, відображаються без будь-якої систематизації та впорядкування. Це зумовлює введення поняття ваги терміна в документі, а в ширшому застосуванні – рейтинг кожного документа в масиві.

Найпростішим методом встановлення ваги терміна в документі є визначення кількості входжень слова в документ (абсолютна частота) та відносної частоти, з якою він трапляється в тексті ($tf_{t,d}$, де *t*, *d* –термін і документ відповідно). Для більш точного визначення ваги терміна використовується поняття частоти документа (df_t) – кількість документів, в яких виявлено даний термін. Зважаючи на те, що при визначенні значущих термінів необхідно дотримуватися таких вимог, як: 1) важливими є терміни, які зустрічаються часто в невеликій кількості документів; 2) менш важливими є терміни, що трапляються менше в документі або більше у великій кількості документів; 3) неважливими є терміни, що виявлені майже у всіх документах, – варто використовувати інвертовану частоту документа:

$$idf_t = \log\frac{N}{df_t},$$

де N – загальна кількість документів у колекції. Відповідно нормовану важливість терміна в документі можна визначити за формулою

$$tf\text{-}idf_{t,d} = tf_{t,d} \times i\ .$$

Виходячи з існуючого поняття ваги терміна, можна скласти певне уявлення про документ, котрий містить набір термінів з певними вагами – важливість (рейтинг) документа, що обчислюється як сума ваг термінів, наявних у ньому:

$$score(q,d) = \sum_{t \in q} tf\text{-}idf_{t,d}.$$

Для порівняння текстових документів між собою, їх каталогізації та загалом більшості завдань, виконуваних ІПС, користуються векторною моделлю подання документа в каталозі. Масив документів у даному випадку є множиною векторів, розташованих в єдиному векторному просторі, де кожна вісь відповідає одному терміну. Для визначення схожості документів між собою користуються методом обчислення косинуса кута між векторами документів:

$$sim(q,d) = \frac{\sum_{i=1}^{M} \vec{V}_i(d_1) \times \vec{V}_i(d_2)}{\sqrt{\sum_{i=1}^{M} (\vec{V}_i(d_1))^2} \times \sqrt{\sum_{i=1}^{M} (\vec{V}_i(d2))^2}}.$$

Використовуючи як один з документів пошуковий запит (q), можна визначити його рейтинг для даного запиту в ІПС:

$$score(q,d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{\left|\vec{V}(q)\right\| \vec{V}(d)\right|}.$$

Завдяки векторній моделі представлення документів є можливість автоматизовано здійснювати каталогізацію наукових текстів без втручання експертів. Хоча при такому способі використання векторної моделі необхідно виконати попереднє «навчання» даної діючої системи на прикладах класифікаційних груп текстів. Для цього в кожну тематичну групу включається достатня для аналізу кількість текстів, з яких формується статистичний портрет на основі векторної моделі. У подальшому класифікатор зможе в автоматичному порядку порівнювати векторне подання існуючих текстів з тими, що надходять у систему.

Проблема пошуку в наукових текстах зводиться не тільки до виділення ключових слів як основних семантично активних елементів тексту, а й до отримання їх рангів (важливостей, ваг). Це дозволяє більш точно визначати тематику тексту та виявляти його семантику. Векторна модель у статистичному аналізі текстової інформації дає змогу отримати базове уявлення про текст на основі припущення, що термін характеризує тематику тексту тим повніше, чим частіше зустрічається в тексті. Проте в даному випадку додатково вводиться поняття фільтрації та нормалізації для відсікання «зайвої» інформації, що не має семантичної активності в документі. Важливою умовою є непорушність семантики документа після проведення даних процедур [11].

Оскільки при побудові векторної моделі документа за координати приймаються частотні показники термінів, наявні в ньому, виникає проблема розмірності аналізованого простору, через те що обсяг словника термінів для кожного документа (якщо не брати до уваги пошукові запити) досягає значних розмірів. Тому досить часто при побудові векторної моделі за координати при-

ймаються не лише частотність окремих термінів-лексем, а й характеристики певних лексемних об'єднань [8]. Хоча варто зазначити, що для наукових текстів після фільтрації в списку термінів розмірність простору слід скорочувати до 5–10 термінів, які спроможні виразити семантику документа. При цьому можна вважати, що збільшення чи зменшення обсягу документа загалом не вплине на кількість семантично активних термінів. Це, у свою чергу, зменшує вимоги до обчислювальних засобів, використовуваних у ІПС.

Семантично активними термінами можуть бути не лише слова, а й словосполучення та інші лексичні одиниці. При розгляді тексту документа як масиву слів постає проблема знаходження їх послідовностей, семантично об'єднаних між собою. Для прикладу, словосполучення «ключове слово» розділити на два слова не можна, бо втрачається смислове навантаження, виражене даною послідовністю слів. На сьогодні існує досить багато методів виділення термінів, що складаються з декількох слів. Переважно всі вони зводяться до автоматичного формування списку словосполучень-термінів і подальшого ранжування стосовно ваги терміна. Досить часто вагу терміна визначають як частоту використання в одному контексті або ж як поняття «mutual information». Для терміна, що складається з двох слів, можна застосувати формулу

$$MI(a) = \frac{freq(a)}{N} * \frac{freq(a)}{0.5 * (Fleft + Fright)},$$

де $freq(a)$ – кількість знайдених сполучень пари a; N–кількість пар; $Fleft$– частота вживання першого слова з терміна як окремого слова; $Fright$ – частота вживання другого терміна з пари. При цьому може вводитись і функція максимальної правдоподібності:

$$\begin{aligned}\log like = {} & a * \log(a+1) + b * \log(b+1) + c * \log(c+1) + d * \log(d+1) \\ & - (a+b) * \log(a+b+1) - (a+c) * \log(a+c+1) \\ & - (b+d) * \log(b+d+1) - (c+d) * \log(c+d+1) \\ & + (a+b+c+d) * \log(a+b+c+d+1),\end{aligned}$$

де a – кількість виявлених сполучень пари; b – кількість виявлених сполучень першого слова з іншими термінами; c – кількість виявлених сполучень другого слова з іншими термінами; d – кількість пар, відмінних від a,b,c.

Використовується також метод C-Value, який збільшує ймовірність визначення словосполучення як терміна при умові, що дане сполучення слів не входить до складу інших сполучень.

$$C\text{-}Value\ (a) = \begin{cases} \log_2|a| * freq(a), \text{якщо сполучення слів не вкладене} \\ \log_2|a| - \dfrac{1}{P(T_a)} * \sum b \in T_a\ freq\ (b) \end{cases},$$

де a – сполучення слів; $|a|$ – кількість слів у сполученні; $freq(a)$ – частотність a; T_a–множина сполучень слів, що містять a; $P(T_a)$ – кількість сполучень слів, що містять a.

Для виявлення поєднання слів у словосполучення використовується також алгоритм, що базується на запам'ятовуванні сусідніх слів із заданим кроком відносно якогось терміна в документі. Безпосередні семантичні складові можуть знаходитись і через покроковий поділ тексту до одиниць словника [2].

При користуванні векторною моделлю аналізу документів необхідно розуміти, що в наборі текстів наявні різноманітні словосполучення, котрі розглядатимуться ІПС як єдиний термін і міститимуть однакові слова, значення яких доволі часто різне, залежить від словосполучення, в яке вони входять [3].

У векторному просторі формується векторне подання слів та інших компонентів текстів шляхом автоматичного видобування статистики їх спільної появи з великих масивів текстової інформації. Ця інформація фіксується в так званих семантичних або контекстних векторах, схожість яких відображає міру семантичної близькості слів [6].

Відтворення семантики текстів в ІПС

Користувачі ІПС при формуванні пошукового запиту досить часто не можуть його точно сформулювати. Проте, навіть коли пошукова система видає на запит користувача результат, то він повинен здійснити аналіз його за релевантністю. Складність формулювання пошукових запитів й отримання релевантних результатів пов'язані також з досить великими лінгвістичними варіаціями, що здатні виражати ту саму думку.

У більшості векторних моделей документ подається як частотний спектр слів, а відповідно і вектор у лексичному просторі. У процесі пошуку близькість між документами можна розглядати як абсолютну, так і відносну. Як абсолютне значення відстані між документами у векторному просторі може бути використане поняття евклідової відстані між координатами документів. Проте такий метод є недосконалим, адже навіть документи, що не мають спільних атрибутів, за даним методом можуть опинитися на ближчій відстані, ніж об'єкти зі спільними атрибутами. Міра близькості між двома об'єктами у векторному просторі буде більш правдоподібною при використанні коефіцієнта Жаккарда. Щоб підрахувати за ним близькість двох об'єктів, необхідно зобразити всі їхні атрибути як точки двох множин. Ці множини перетинаються в точках, де об'єкти мають спільні атрибути. Коефіцієнт Жаккарда обчислюється як частка перетину даних множин до їх об'єднання [10]. Проте більш застосовуваним способом виявлення схожості документів є функція косинуса кута, що утворюють між собою документи у векторному просторі. Такий метод є відносним, оскільки не прив'язаний до розмірності векторного простору й дозволяє здійснювати ранжування і пошук семантично близьких документів [1].

Варто зазначити, що суть векторної моделі подання текстів полягає в статистичному виділенні ключових слів, кількість яких під час первинного аналізу в десятки разів перевищує кількість текстів, у котрих вони містяться. При роботі з даною моделлю можуть виникати проблеми –при виділенні ключових слів зазвичай не враховуються форми слова, а сам процес здійснюється з використанням чітких алгоритмів без урахування невизначеностей. До

того ж виділені ключові слова не можна в повній мірі вважати ознаками тексту, оскільки текст – це не лише множина слів, а й структура з певним внутрішнім порядком [3].

Якщо розглядати тексти наукового спрямування, то слід звернути увагу на те, що такі документи мають певну структуру заголовків й елементів форматування, вміщають найбільш значущі слова. Заголовки зазвичай також мають ієрархію (підзаголовки), а слова в них – різну вагу [10]. При цьому завдання формування списку ключових термінів спрощується, оскільки структура заголовків виступає по суті в ролі семантичної структури документів. Єдиним нюансом може бути те, що автор не зумів коректно вибрати назву заголовка, а це впливатиме на сприйняття семантики тексту. При аналізі документів доцільно для заголовків застосовувати ті самі процеси, що й для іншого тексту, з урахуванням коефіцієнта рівня заголовка, в якому зустрічається слово (наприклад, при підрахунку частоти вживання слова).

Використовуючи статистичний аналіз текстів, з урахуванням моделі подання даних, можна створювати професійні системи і бази знань, здійснювати узагальнення інформації та реферування текстів документів, формувати онтології в тій чи іншій професійній сфері, виконувати пошук і фільтрацію текстових документів [1].

В існуючих програмних реалізаціях ІПС при аналізі текстів виконують наступні операції: індексування корпусу, фільтрація словника, побудова матриць частот, перетворення, зважування та нормування матриць, побудова контекстних векторів, застосування контекстних векторів для формування представлень документів [6].

Ураховуючи те, що основним завданням при роботі з повнотекстовими базами даних є пошук документа за його вмістом [2], векторна модель здатна якомога повніше описати семантику кожного документа і колекції загалом. На основі частотних статистичних показників можна сформувати базис семантичного простору текстових документів. Ієрархічна кластеризація документів у такому просторі дає можливість згрупувати між собою тематично близькі документи [9]. Математичні моделі подання текстів забезпечують проведення різноманітних статистичних досліджень, зокрема побудову статистичного портрета із заданими параметрами та його використання для виявлення мови документа і теми тексту, порівняння текстів між собою.

Таким чином, метод векторного подання текстової інформації дозволяє здійснити семантичний аналіз тексту на основі статистичних показників і виділити в ньому семантично активні елементи й структурно-лінгвістичні зв’язки між ними.

1. Векторная модель представления текстовой информации : материалы Междунар. науч. конф. “Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам”. – Ижевск, 2006. – 131 с. 2. Войтиков В.А. Один метод, позволяющий оптимизировать содержание курса лекций / В.А. Войтиков, В.Ф. Пугач // Вісн. Східноукраїн. нац. ун-ту ім. В. Даля. – 2008. – №9 (127), Ч.2. – С. 33–41. 3. Дідковська М.В.

Статистична модель аналізу текстів / М.В. Дідковська, Д.В. Старосуд // Нові технології. – 2011. – №2 (32) – С. 62 – 69. 4. К вопросу о методах анализа документов в информационной деятельности / Н.В. Махортова, И.А. Омельяненко, С.Г. Шапошникова и др. // Вісн. Східноукраїн. нац. ун-ту ім. В. Даля. – 2008. – №8 (126), Ч.1. – С. 359–362. 5. Крыгин М.Ю. Текст на естественном языке как объект статистического анализа / М.Ю. Крыгин // Бионика интеллекта. – 2010. – №1(72). – С. 75 – 82. 6. Мисуно И.С. Векторные и распределенные представления, отражающие меру семантической связи слов /Мисуно И.С., Рачковский Д.А., Слипченко С.В. // Математичні машини і системи. – 2005. – №3 – С. 50–66. 7. Об одном методе статистической фильтрации текстовой информации: материалы Междунар. науч. конф. “Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам”. – Ижевск, 2006. – 126 с. 8. Павлишенко Б. Семантична кластеризація текстових документів методом k-середніх / Б. Павлишенко // Вісн. Нац. ун-ту «Львівська політехніка». –2011. – № 710. – С. 215–218. 9. Павлишенко Б. Групування текстових даних на основі моделі семантичного контексту / Б. Павлишенко // Восточно-Европ. журн. передовых технологий. – 2011. – №5/2 (53). – С. 39 – 42. 10. Текстовая кластеризация алгоритмом ROCK : материалы XVII всероссийской науч.-методич. конф. [«Телематика’2010»]. [Электронный ресурс] / Савин. И.И. – 2010 – Режим доступа: http://tm.ifmo.ru/tm2010/src/263e.pdf. 11. Christopher D. Manning. Introduction to information retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze // Cambridge University Press. – Cambridge, 2008.

ПРЕДСТАВЛЕНИЕ СЕМАНТИКИ ТЕКСТА ПОСРЕДСТВОМ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Проанализированы методы представления текстовых документов в информационно-поисковых системах. Описаны возможности представления содержания текста посредством статистических показателей.

PRESENTATION OF TEXT SEMANTIC THROUGH ITS STATISTICAL INDICATORS

It was analyzed methods represent text documents in information retrieval systems and the possibility of meaning representation of text through its statistics.

Стаття надійшла 14.11.2012

УДК 004

М. Козелко

Українська академія друкарства

МОДЕЛЬ ФУНКЦІОНУВАННЯ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ РЕДАКТОРА ГРАФІЧНИХ УНІТЕРМІВ

Засобами алгебри алгоритмів описано модель функціонування інструментальних засобів редактора графічних унітермів.

Унітерм, абстракція, графічний інтерфейс, модель, система, проект