УДК 004.421 (045)

Заболотня Т.М.,
Сапсай Т.Г., канд. техн. наук

РЕАЛІЗАЦІЯ АВТОМАТИЗОВАНОЇ ОРФОКОРЕКЦІЇ ТЕКСТОВИХ ДАНИХ В ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ

Національний технічний університет України «КПІ»

Обґрунтовано необхідність введення орфокоректора до складу інформаційно-пошукової системи з метою забезпечення її коректної роботи зі спотвореними текстовими даними. Визначено зміни у структурно-алгоритмічній організації ІПС, які може спричинити підключення до неї коректора. Показаний вплив характеристик ІПС на будову коректора у разі, коли він створений у формі програмного агента

Вступ

На сьогоднішній день інформаційно-пошукові системи (ІПС) є одним з найбільш затребуваних класів програмного забезпечення. З огляду на постійне збільшення обсягу електронної документації, проблема підвищення ступеня автоматизації обробки даних такими системами є актуальною завжди.

Важливою умовою забезпечення високого рівня автоматизації ІПС є здатність останніх до коректного функціонування за умови роботи із спотвореними текстовими даними. Існуючі програмні рішення здебільшого не мають окремого модуля орфокорекції, а реалізують алгоритми квазівиправлення помилок: пропонують до подальшої обробки найчастіше вживаний формально вірний варіант виправлення спотвореного слова [1, 2]. Це спричиняє необхідність залучення користувача до процесу визначення вірних варіантів написання слова з помилкою і, відповідно, знижує рівень автоматизації функціонування ІПС.

Виходячи з вищезазначеного, метою даної статті є дослідження доцільності введення до структури інформаційнопошукової системи програмного орфокоректора, визначення особливостей реалізації такого орфокоректора в залежност від характеристик ІПС, а також аналіз змін у структурно-алгоритмічній організації ІПС, які можуть бути цим викликані.

Місце орфокоректора в інформаційно-пошуковій системі

Проаналізуємо доцільність введення орфокоректора до складу інформаційно-пошукової системи, яка має типову для більшості ІПС структурно-алгоритмічну організацію [2, 3]. Коректор непотрібний для підтримки всіх без винятку функцій системи, тому визначимо режими роботи останньої, у яких може виникати необхідність у виправленні орфографічних помилок, а також складові системи, які при цьому можуть бути задіяними.

Можна виділити три функціональних профілі інформаційно-пошукової системи: користувача, адміністратора та оператора.

Користувачу інформаційнопошукова система може надавати такі послуги як [2]:

- атрибутивний пошук документів;
- повнотекстовий пошук документів;
- пошук документів за тезаурусом;
- каталогізація документів;
- кластеризація обраних документів.

Атрибутивний пошук документів передбачає їх фільтрацію за значеннями атрибутів таких, наприклад, як автор та назва документа, дата його створення тощо. Більшість цих даних мають чисельний формат, є датами або рядками, що містять одне єдине слово, тому застосування в такій ситуації орфокоректора неможливе.

Під *повнотекстовим пошуком* розуміється фільтрація документів за наявністю в них певних слів або фраз. У випадку, коли в рядку пошуку одне чи декілька слів є спотвореними, обробка запиту системою не

може бути виконана, адже за словоформою з помилкою не можна встановити її базової форми, а значить, неможливо здійснювати подальший повноцінний пошук документів. Тому необхідно забезпечити підтримку даного режиму роботи пошукової системи функцією орфокорекції. При цьому потрібно звернути увагу на кількість слів у запиті користувача. Якщо у рядку пошуку буде введене одне спотворене слово, орфокоректор, який реалізує контекстноорієнтований метод виправлення помилок [4], буде нездатний сформувати необхідні варіанти виправлення. У такому випадку він зможе визначити близькі до заданого слова словоформи тільки за формальними ознаками їх схожості.

Обґрунтувати доцільність виправлення орфографічних помилок у запиті користувача при роботі інформаційнопошукової системи в режимі повнотекстового пошуку можна з декількох позицій. По-перше, корекція слів у такому випадку не потребує синтаксичного узгодження варіантів виправлення, адже для ІПС важливим є визначення базової форми слова [2, 3]. По-друге, у випадку реалізації орфокоректором контекстноорієнтованого методу виправлення помилок у ролі контексту можуть виступати всі слова запиту. Відносно невелика кількість слів у запитах (~ 71% запитів складаються з 2-4 слів [2]) не є перешкодою для застосування і семантичного аналізу, тому що навіть одне вірно написане ключове слово може визначити область пошуку варіантів виправлення. По-третє, користувач під час складання запиту до ІПС намагається використовувати ключові слова, які найбільш адекватно відображають його інформаційну потребу та є максимально семантично навантаженими. Тому імовірністшвидкої та точної обробки пошукових запитів є високою.

Таким чином, алгоритм повнотекстового пошуку доцільно доповнити етапом орфокорекції невірно введених слів.

1. Розбиття запиту на окремі слова.

2. Перевірка наявності кожного слова у словнику та пошук його лексеми.

3. Якщо слова, яке аналізується, у словнику немає, вважається, що у ньому є помилка, та робиться спроба її виправити.

4. У разі знаходження варіантів виправлення спотвореного слова, визначаються їх базові словоформи, відносно яких далі ведеться пошук у індексному сегменті.

5. У разі, коли орфокоректор на основі переданих до нього даних не визначив варіанти виправлення спотвореного слова, пошук в базі ведеться відносно вихідного написання цього слова.

Пошук документів за тезаурусом передбачає фільтрацію документів у відповідності до наявності в них ключових слів, які належать обраній гілці тезаурусу. Для проведення пошуку такого виду користувач повинен вибрати у вже сформованому тезаурусі гілку, відносно термінів якої буде виконуватися пошук документів. Тезаурусний ресурс на може містити помилки у назвах вузлів, тому застосування орфокоректора у даному режимі роботи інформаційнопошукової системи є непотрібним.

Каталогізація передбачає ієрархічне групування документів за обраними ознаками, у тому числі за гілками наявних в системі класифікаторів. Генерація дерева каталогу відбувається на основі списку спільних атрибутів документів, які відносяться до однієї гілки. Це означає, що, як і у випадку атрибутивного пошуку, під час каталогізації немає необхідності у звертанні до програмних засобів орфокорекції.

Кластеризація ряду обраних документів, тобто виділення в них кластерів, які подібні за змістом, може потребувати виконання орфокорекції, тому що групи формуються на основі попарної схожості текстів документів, в словах яких можуть бути присутні помилки. Можливим застосування орфокоректору, який реалізує контекстноорієнтований метод виправлення спотворень, в даному випадку робить наявність оточення слова з помилкою. Місце орфокорекції у алгоритмі кластеризації документів є подібним до його місця в алгоритмі повнотекстового пошуку: у випадку, коли слово, яке аналізується, не належить словнику,

робиться спроба знайти варіанти його виправлення та продовжувати обробку документа з урахуванням останніх.

Отже, у межах функціонального профілю користувача ІПС проведення орфокорекції може бути потрібним для виконання повнотекстового пошуку документів, а також кластеризації їх певної відібраної множини.

Оператор інформаційно-пошукової системи може здійснювати:

- внесення документу до системи;
- перегляд атрибутів внесеного документу та перевірку правильності автоматичної класифікації.

Процес *додавання документів* до системи складається з декількох кроків, найважливішим з яких є генерація інвертованого індексу за текстом документу [2], адже пошук проводиться саме за індексними даними. Розглянемо алгоритм індексації та визначимо місце орфокорекції у ньому.

Послідовність дій щодо побудови індексів є такою [2]:

- поділ тексту документа на слова;
- пошук базової словоформи для кожного слова, а також індексу останньої у лінгвістичній базі;
- визначення номеру кожного слова в документі для реалізації можливості пошуку за відстанню між словами;
- формування для кожного слова характеристичного рядка;
- внесення отриманих характеристичних рядків до бази.

Необхідність у звертанні до орфокоректора може виникнути на етапі пошуку у словнику лексем (базових словоформ) заданих слів. У випадку, якщо задане слово не міститься у словниковій БД, воно може бути спотвореним, а, значить, індексація документу за цим словом буде безрезультатною. Слід зазначити, що ключові слова, які несуть семантичне навантаження, в тексті документа зустрічаються не так часто (а іноді – тільки в назві документа) [5]. Тому індексація документа за невірним написанням таких слів може привести до того, що даний документ взагалі не буде знайдений інформаційнопошуковою системою у разі введення правильно написаних ключових слів до рядку запиту.

З огляду на вищезазначене пропонується ввести крок щодо підбору вірного

варіанту написання спотвореного слова до алгоритму генерації інвертованих індексів. Це дозволить на основі даних про оточення слова, яке не міститься у словнику, знайти вірний варіант його написання і занести його до індексної бази, таким чином забезпечивши знаходження відповідного документу за цим ключовим словом.

Перегляд атрибутів доданого до системи документу та *перевірка результатів класифікації* не потребують втручання з боку програмних засобів орфокорекції, адже ці дії мають виключно наглядовий характер.

Адміністратор системи має можливість виконувати такі дії як:

- регламентування доступу до ресурсів та режимів системи;
- створення нових типів атрибутів;
- створення нових типів документів;
- створення та редагування довідників,
- жодна з яких, як правило, не передбачає проведення виявлення та виправлення орфографічних помилок.

Отже, в результаті аналізу функціональних профілів інформаційнопошукової системи виявлено режими, під час роботи в яких може виникнути необхідність у виправленні спотворених слів: це режими повнотекстового пошуку документів, їх кластеризації, а також режим генерації інвертованого індексу при внесенні нових документів до системи.

Визначимо тепер місце орфокоректора у структурі ІПС. Інформаційнопошукові системи, як правило, орієнтовані на використання у *Web*-середовищі [1], та складаються з таких частин як:

- підсистема клієнта, яка забезпечує інтерактивний інтерфейс користувача;
- підсистема сервера, що забезпечує трансляцію клієнтських запитів в запити до бази даних;
- інформаційна база даних.

Рис. 1. Структурна схема інформаційно-пошукової системи

До задач *клієнтської підсистеми* входить реагування на дії користувача та відправлення запитів до серверної підсистеми. Як зазначено вище, звертання до орфокоректора може мати місце у режимі повнотекстового пошуку. Тому можна зробити припущення щодо доцільності його включення до складу клієнтської підсистеми: одразу після введення запиту користувача виконувати перевірку наявності у ньому спотворених слів, виправляти їх і вже тоді відправляти на обробку серверній частині. Але орфокоректор у своїй роботі може використовувати словникові ресурси [4, 6], які мають бути розташовані на одному з ним вузлі комп'ютерної мережі, з метою уникнення необхідності передавання великих обсягів даних каналом зв'язку. Якщо задовольнити цю вимогу, необхідно кожний клієнт ІПС забезпечити відповідними словниковими ресурсами, що значно збільшить його ресурсоємність. Отже, включення орфокоректора до складу клієнтської підсистеми є недоцільним.

Серверна підсистема містить модулі, пов'язані із визначеними вище режимами роботи ІПС, у межах яких може виникати необхідність у звертанні до орфокоректора (рис. 1). Крім того, інформаційна

база ІПС, як правило, розміщується на тому ж комп'ютерному вузлі, що й серверна підсистема. Тому виправданим є введення орфокоректора до складу даної підсистеми та налаштування його на обробку повідомлень від модуля повнотекстового пошуку, кластеризатора, класифікатора та індексатора. Словникові ресурси, необхідні для роботи орфокоректора, у такому випадку входитимуть до інформаційної бази.

Аналіз впливу характеристик ІПС на структурно-алгоритмічну організацію модуля орфокорекції

Для використання у складі інформаційно-пошукової системи орфокоректор доцільно реалізовувати на основі агентоорієнтованого підходу до створення програмного забезпечення [3, 7]. Підґрунтям для цього є той факт, що на сьогоднішній день саме програмні системи пошукового типу найчастіше мають відкриту архітектуру, причому багато з них представлені у формі багатоагентної системи. У [6] наводяться результати дослідження впливу зовнішнього середовища (абстрактної системи АОТ) на структуру та поведінку агента-орфокоректора. Проаналізуємо, яким чином характеристики інформаційно-пошукової системи як зовнішнього середовища можна врахувати при розробці програмного агента-орфокоректора.

Структура інформаційно-пошукової системи. Інформаційно-пошукова система, яка розглядається у даній статті, має віддалене розміщення компонентів. Але всі модулі серверної підсистеми, під час роботи яких може виникнути потреба у проведенні орфокорекції текстових даних, розташовані на одному вузлі комп'ютерної мережі, тому орфокоректор достатньо побудувати у формі реактивного програмного агента.

Ресурси, які потрібні для роботи агенту-орфокоректору, у інформаційній базі ІПС можуть бути представлені неповністю: часто у наявності є тільки лексикографічний словниковий ресурс або частотний словник [1]. У такому разі до множини баз даних, які використовує у своїй роботі інформаційно-пошукова система, слід додати лексико-семантичний словник (наприклад, популярний словниковий ресурс *WordNet,* розроблений ученими Прінстонського університету [8, 9]). Слід зазначити, що такий ресурс може бути корисним не тільки у процесі визначення варіантів виправлення спотворених слів, але і під час розв'язання будь-яких інших задач, пов'язаних з семантичною обробкою текстових даних (наприклад, при пошуці документів за тезаурусом або у процесі кластеризації документів тощо).

Функціональні характеристики ІПС. Інформаційно-пошукова система під час обробки текстових даних оперує здебільшого базовими формами слів (лексемами). Тому відсутня потреба у синтаксичній узгодженості варіантів виправлення, які підбирає програмний орфокоректор, із контекстним оточенням спотвореного слова.

У разі, якщо *текстові дані*, з якими працює інформаційно-пошукова система, належать до лексики з певної предметної галузі, орфокоректор можна налаштувати на роботу з ними шляхом модифікації вмісту відповідного лексикосемантичного словника [8].

Пріоритетність критеріїв оцінювання ефективності роботи інформаційнопошукової системи в режимах, де може виникнути потреба у звертанні до орфокоректора, є різною. Під час повнотекстового пошуку перевага надається швидкості підбору варіантів виправлення спотворених ключових слів у запиті, оскільки даний тип пошуку виконується за безпосередньою участю користувача, котрий може вибрати вірний варіант написання слова з-поміж тих, які запропоновані системою.Для кластеризації та індексації документів найважливішим показником є точність виправлення помилок. Дані операції здійснюються автоматично, часто – тривалий проміжок часу, тому одразу виправити допущені орфокоректором неточності неможливо. Усі похибки у виправленні орфографічних помилок в текстових даних будуть доступні для аналізу лише по завершенню процесу індексації чи кластеризації. Кінцевий результат кластеризації сегменті інформаційної бази

пошукової системи внаслідок індексації нових документів відстежити неможливо. Тому для підтримки роботи системи в режимі генерації індексів коректор має реалізовувати контекстноорієнтований метод виправлення орфографічних помилок [4].

Програмно-апаратне забезпечення інформаційно-пошукової системи. Вище визначено, що доцільним місцем орфокоректора в структурі ІПС є серверна підсистема. Тому для перевірки можливості ефективної роботи орфокоректора в складі інформаційно-пошукової системи необхідно звертати увагу на характеристики програмно-апаратного забезпечення саме цієї підсистеми.

Висновки

Показано доцільність введення орфокоректора до складу інформаційнопошукової системи. Проаналізовано функціональні профілі ІПС та визначено, що виправлення помилок у спотворених словах може бути необхідним у процесі повнотекстового пошуку документів, їх кластеризації та генерації інвертованого індексу документів. Для того, щоб модулі пошукової системи могли ініціювати роботу орфокоректора, його запропоновано розмістити на серверному боці ІПС як частину модуля лінгвістичної підтримки. Визначений вплив характеристик інформаційно-пошукової системи на структурно-алгоритмічну організацію коректора, побудованого у формі програмного агента.

Більш детального вивчення потребує питання сумісного використання електронних лінгвістичних ресурсів інформаційно-пошуковою системою та коректором, який входить до її складу. Також перспективним напрямом подальшого дослідження проблеми забезпечення коректної обробки пошуковою системою спотворених текстових даних є розробка протоколів взаємодії програмного орфокоректора та інших модулів ІПС.

Список літератури

1. *Белозеров В.Н., Кулькова Г.В.* Лингвистическое обеспечение корпоративной информационно-поисковой системы // НТИ, сер.1. – 2004. – №3. – С. 14–18.

2. *Ландэ Д.В.* Поиск знаний в Internet. – М.: Диалектика, 2005. – 271 с.

3. *Плескач В.Л., Рогушина Ю.В.* Агентні технології: Монографія. – К.: Київ. нац. торг.-екон. ун-т, 2005. – 338 с.

4. *Михайлюк А.Ю., Заболотня Т.М.* Комбінований метод виправлення орфографічних помилок у текстових даних // Вісник Хмельницького національного університету. – 2007. – № 2. – Т.2. – С. 21–26.

5. *Якушин Б.В.* Алгоритмическое индексирование в информационных системах. Проблематика и методы. – М.: Наука, 1978. – 144 с.

6. *Заболотня Т.М., Михайлюк А.Ю, Тарасенко В.П.* Структурно-алгоритмічна організація програмного агентаорфокоректора // Наукові вісті НТУУ "КПІ". – 2008. – №1. – С. 88–96.

7. *Бугайченко Д.Ю., Соловьев И.П.* Абстрактная архитектура интеллектуального агента и методы её реализации // Системное программирование. – СПб.: СПбГУ" – 2005. – № 1. – С. 36–67.

8. *Заболотня Т.М.* Селективний підхід до автоматизованого формування оперативного спеціалізованого лексикосемантичного словникового ресурсу // Бионика интеллекта. – 2007. – № 2(67). – С. 27–31.

9. *Miller G.A.* WordNet: a Lexical Database for English // Communications of the ACM. – 1995. – Vol.38. – №11. – P. 39–41.