

Стаднікова А.Є., магістрант, Вербицький В.Г., професор, науковий керівник,
Безверхий А.І., доцент, консультант

ДОСЛІДЖЕННЯ ПРОБЛЕМИ НЕЧІТКОГО ПОШУКУ У ТЕКСТІ ТА СЛОВНИКУ НА ОСНОВІ МЕТОДУ ОПОРНИХ ВЕКТОРІВ

Запорізька державна інженерна академія, кафедра ПЗАС

З кожним роком обсяги Інтернету збільшуються в рази, тому ймовірність знайти необхідну інформацію різко зростає. Інтернет об'єднує мільйони комп'ютерів, безліч різних мереж. Число користувачів збільшується на 15-80% щорічно. І, тим не менш, все частіше при зверненні до Інтернет основною проблемою виявляється не відсутність шуканої інформації, а неможливість її знайти. Як правило, звичайна людина в силу різних обставин не може або не хоче витратити на пошук потрібної їй відповіді більше 15-20 хвилин. Саме тому надзвичайно важливим є питання формування орфографічно правильних запитів, адже у разі помилки користувач може не отримати необхідних даних або отримати помилкові дані.

Задача аналізу алгоритмів нечіткого пошуку на даний момент актуальна, так як область застосування таких алгоритмів неймовірно велика і різноманітна. Деякі з них успішно застосовуються в біоінформатиці для порівняння генів, білків, хромосом, для роботи з базами даних в системах моніторингу лісопожежної ситуації, у розпізнаванні рукописних символів на пристроях з сенсорним екраном та багатьох інших сферах. Найбільш поширеною областю застосування на сьогоднішній день є функції на зразок «Можливо ви мали на увазі ...» у пошукових системах Google, Yahoo та ін., які використовуються для виявлення помилок в полях введення програм. Однак, подібні існуючі системи не надають можливість вільного доступу до свого функціоналу та його використання. Саме тому є доцільним створення додатку для обробки текстових даних та автоматичного виявлення помилкових слів, який може бути використаний у роботі пошукових систем або на звичайних веб-сайтах.

Для досягнення необхідного результату використовуються:

- Метрики відстані Левенштейна [4].
- Метод опорних векторів (Support Vector Machine, SVM) [1,3].
- Семантичний словник WordNet

У якості семантичного словника було обрано мережу WordNet. Вибір було зроблено на основі наступних особливостей:

- WordNet об'єднує слова в набори синонімів, які називаються синсети, надає короткі визначення та приклади використання, також записує ряд відносин між цими синонімами або їх членами
- База даних містить більше 155 287 слів, організованих у 117 659 синсетах
- WordNet можна вільно використовувати в комерційних і наукових цілях

Відстань Левенштейна дозволяє розширити область пошуку, вказуючи відстань до ключових слів, яка є мінімальною кількістю операцій вставки, видалення і заміни, необхідних для перетворення однієї послідовності символів в іншу.

Для написання додатку було обрано мову програмування PHP, базоданих MySQL та програмне середовище PhpStorm.

Розроблений програмний комплекс складається з наступних модулів:

- **Модуль обробки вхідних даних**

Після введення вхідних даних необхідно розбити текст на окремі слова та виключити усі орфографічні знаки, а також усі лексичні категорії, окрім іменників, дієслів, прикметників та прислівників для подальшої обробки.

- **Модуль визначення метрики**

Для визначення метрики використовується відстань Левенштейна, яка розраховується за допомогою генерації матриці, враховуючи вартість операцій вилучення, заміни та вставки символів.

- **Модуль пошуку даних у семантичній мережі**

По заданому слову, використовуючи семантичний словник, знаходяться усі слова, які співпадають з даним словом (або починаються з нього) та обирається найбільш близьке за значенням.

Дослідження, проведені за допомогою розробленого програмного додатку показали, що сфера нечіткого пошуку у тексті та словнику займає важливе місце у роботі сучасних пошукових систем, тому дослідження у цій сфері мають практичний сенс, оскільки в даний момент на ринку відсутні прикладні рішення для вільного використання функціоналу виявлення орфографічних помилок у тексті.

Висновки:

- Проведене дослідження проблеми нечіткого пошуку, методів та алгоритмів її вирішення.
- Досліджені методи обробки текстових даних та маніпуляцій з ними.
- Досліджена архітектура та принцип побудови семантичної мережі WordNet .
- Встановлено актуальність розробки даного програмного засобу.

Література

1. *Alexey Nefedov*. Support Vector Machines: A Simple Tutorial. — 2016.
2. *Владимир Вьюгин*. Математические основы машинного обучения и прогнозирования. — МЦМНО, 2014. — 304 с. — ISBN 978-5-457-71889-0.
3. *Nello Cristianini, John Shawe-Taylor*. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. — Cambridge University Press, 2000. — ISBN 978-1-139-64363-4.
4. В. И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965. 163.4:845-848.