

ДОСЛІДЖЕННЯ МОДЕЛЕЙ ТА ПРОГРАМНИХ КОМПОНЕНТІВ ДЛЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ ПОШУКУ РОБОТИ

Рудя Е.В., Лютенко І.В.

***Національний технічний університет
«Харківський політехнічний інститут», м. Харків***

Розповсюдженою проблемою є недостатня якість існуючого програмного забезпечення з процесу пошуку роботи. Бажано мати програмні компоненти, які б були інтегровані в одну систему, проводили пошук інформації та аналіз даних відповідно до вимог суб'єктів. У роботі проводиться аналіз та розробка інструментів (API парсерів) для інтеграції програмних рішень та порівняльний аналіз даних при процесі пошуку вакансій, які зібрані з різних джерел. Реалізовані методи автоматичного поділу web-сторінок на службово-навігаційну та семантично-змістовну частини. Гіпотеза представляє собою припущення корисності виключення та зменшення ваги службової частини html-сторінок з індексу пошукової системи з метою реалізації завдання підвищення якості інформаційного пошуку. У роботі використаний підхід, що базується на визначенні областей, які повторюються у множині сторінок одного сайту. Запропоновано алгоритм для виконання описаної задачі.

На вхід алгоритму надходить деяка директорія з html-файлами, з відповідними сторінкам одного аналізованого сайту. Алгоритм виконує обробку даних файлів, визначення та маркування в них ділянок програмного коду, які вважаються службово-навігаційними частинами. З різними налаштуваннями алгоритмом проводиться або видалення службово-навігаційних елементів з html-файлів, або маркування (виділення) цих частин спеціальними html-тегами.

Умовно алгоритму можна поділити на кілька етапів.

1 Розподіл html-файлів на базові, неподільні при подальшому порівнянні символно-кодові послідовності-фрагментарів.

2 Виконання пошуку файлових підмножин (кластерів), з набором, що повторюється з ланцюжків фрагментів (кластери визначають деяку підмножину сторінок з єдиною службово-навігаційною частиною). Набір ланцюжків фрагментів представляється одним або кількома ланцюжками фрагментів.

3 Видалення (або маркування) певної послідовності ланцюжків фрагментів зі всіх файлів даного кластера.

4 У разі не покриття кластером усієї множини файлів, необхідно заново виконати попередні кроки для решти множини документів.

5 У разі наявності залишку недоторканих файлів, в цих файлах виконується новий пошук ланцюжків фрагментів з інших кластерів.

Застосування подібних алгоритмів, може показати ефективність для спеціалізованих завдань інформаційного пошуку у випадках, коли є зниження якості пошуку внаслідок зашумлення службово-навігаційними елементами оформлення html-сторінок конкретного сайту.