УДК 519.8

Парсинг у лінгвістиці та інформатиці

***М. О. Лєбєдєва**, студентка*
Полтавський університет економіки і торгівлі
mariialiebiedieva@gmail.com

***О. О. Черненко**, к. ф.-м. н.*
Полтавський університет економіки і торгівлі
oksanachernenko7@gmail.com

В статті розглядається автоматизований метод пошуку інформації – парсинг. Використання методу парсингу для ефективного пошуку інформації.

Liebiedieva M.O. The article describes an automated method of finding information - parsing. Using the Parsing Method to effectively find information.

Ключові слова: КОНТЕКСТНО-ВІЛЬНА ГРАМАТИКА, ПАРСИНГ, ПОШУКОВА ОПТИМІЗАЦІЯ.
Keywords: CONTEXT-FREE GRAMMAR, PARSING, ENGINE OPTIMIZATION.

Нині людство все частіше використовує інформацію в цифровому вигляді. І часто виникає необхідність для конкретного аналізу цієї інформації та її структурування. З цією метою використовується синтаксичний розбір (парсинг) у лінгвістиці та інформатиці – процес порівняння лінійної послідовності лексем (слів) природної або формальної мови з її формальною граматикою. Результатом є, як правило, дерево розбору (синтаксичне дерево).

Під час синтаксичного аналізу текст оформляється у структуру даних, зазвичай – в дерево, яке відповідає синтаксичній структурі вхідної послідовності, і добре

підходить для подальшої обробки. Зазвичай синтаксичні аналізатори працюють в два етапи: на першому ідентифікуються осмислені токени (виконується лексичний аналіз), на другому створюється дерево розбору. Кожна мова програмування має точні правила, які задають синтаксичну структуру коректних програм.

Парсинг сайтів є ефективним рішенням для автоматизації збору інформації, адже ця програма має здатність знаходити серед тисячі сайтів потрібну інформацію, відкидати зайве, упаковувати дані в необхідному вигляді, і в подальшому науковець уже може забирати релевантні дані для їх використання у створенні та розробці своєї наукової роботи. За допомогою нього можна оцінити як швидко програма може впоратись із поставленим завданням, чи достовірна та чи інша інформація, чи взагалі відчувається релевантність у процесі роботи тощо. Метод парсингу має кілька етапів для витягнення потрібної інформації із наукових сайтів:

1. Збір контенту.
2. Витяг інформації.
3. Збереження результатів.

Запропонована методика дасть можливість дослідити та проаналізувати дану інформацію, чи потрібна вона дослідникові для внесення її у свою роботу.

Алгоритм Earley — алгоритм розбору запропонованих даних для контекстно-вільної граматики; він заснований на методі динамічного програмування [2]. Цей алгоритм не накладає ніяких обмежень на аналіз контекстно-вільної граматики, яка використовується. Алгоритм Earley реалізує стратегію проходу «зліва направо». Алгоритм синтаксичного аналізу може бути представлений у вигляді обчислюваної функції розбору з двома аргументами Parse (G, ω):

- G={N,T,P,S} — контекстно-вільна граматика з

великою кількістю не-термінальних символів N, множиною терміналів T, набором правил P і початковою граматикою нетерміналів S;

• ω = a1…an — рядок з n термінальних символів граматики G. Функція розбору Parse повертає множину дерев виведення вхідного рядка ω, якщо вона виведена в граматиці G, і значення False в іншому випадку. Синтаксичний аналізатор Earley здійснює аналіз алгоритму вхідного рядка символів за рахунок проходження знизу догори і отримується єдиноправильний вихід вхідного рядка, якщо вхідна граматика є однозначною, або набір правил виведення, якщо вхідна граматика є неоднозначною. Оригінальний алгоритм Earley тільки виявляє вхідний рядок, але не розбирає його.

Алгоритм Earley використовує три обчислювальні процедури для побудови станів:

• Сканер (Si): сканує кожен елемент у стані Si і, якщо символ Xp дорівнює терміналу ai+1 у деяких ситуаціях[r,p,j], додає до значення стану Si+1.

• Предиктор ([r,p,j],Si): перевіряє, чи є символ Xp нетермінальним символом граматики G, і якщо так, він перевіряє, чи Lr'=Xp виконується для кожного правила r' граматики G, якщо так, то ситуація [r', 0, i] додає до значення стану Si.

• Укладач([r,p,j],Si): сканує кожну ситуацію [r', p', k] станів Sj, і якщо Xp'=Lr, то ситуація [r', p'+1, k] додає до значення стану Si.

Процедуру Сканер (Si) викликають передусім, щоб заповнити стани Si, тоді до нової ситуації [r,p,j] додавання до стану Si, потім викликаються процедури Предикатор([r,p,j],Si) і Укладач([r,p,j],Si): для кожної доданої ситуації.

Отже, підсумовуючи вищесказане, можна зазначити, що використання саме парсингу для підвищення ефективності

пошуку інформації для дослідників є найбільш дієвим способом, особливо, якщо є ключові слова.

Можна відзначити, що деякі алгоритми працюють швидше, ніж інші. Але водночас не всі алгоритми можуть працювати з усіма граматиками. Налаштувавши роботу, можна оперативно підібрати необхідні для просування запити, тому що парсер за короткий термін обходить тисячі сторінок, фільтрує представлені дані, відбираючи серед них потрібні, після чого пакує отриманий результат для подальшої обробки.

Література

1. Давидов М.В., Лозинська О.В., Пасічник В.В. Ефективний алгоритм для синтаксичного аналізу речень з використанням семантично позначених зважених афіксних контекстно-вільних граматик / Давидов М.В., Лозинська О.В., Пасічник В.В. - Радіоелектроніка. Інформатика. Управління. - 2017. - № 4. - С. 124-130.
2. Інтелектуальна обробка текстів: / В.Ю. Тарануха. – Київ: електронна публікація на сайті факультету, 2014. – 80 с.
3. Порівняльний аналіз методів синтаксичного розбору текстів: / І.Б. Швороб.-Національний університет «Львівська політехніка», 2015-197-202 с.