Python 언어 기반의 다중 프로세스를 이용한 채용공고 수집기

허태성*, 김준호O, 백상헌*
인하공업전문대학 컴퓨터정보과O
인하공업전문대학 컴퓨터정보과*
e-mail: tshur@inhatc.ac.kr*, kalmia6863@gmail.comO, kororo1019@gmail.com*

Recruitment collector using multiple processes based on Python

Tai-Sung Hur*, Jun-Ho KimO, Sang-Heong Baek*
Dept. of Computer Information, Inha Technical CollegeO
Dept. of Computer Information, Inha Technical College*

● 요 약 ●

평생직장이 사라지면서 중년층은 재취업을 희망하고, 매년 실업률이 높아지면서 취업 포털 사이트를 이용하는 이용자들이 증가하고 있다. 이러한 이용자들에게 채용공고 정보를 제공해 주기위해서 보편적으로 Python 언어를 이용하여 데이터를 처리하고 수집한다. 하지만, Python은 다중 프로세스 기반을 갖춘 PC가 보급됨에도 불구하고 단일 프로세스로 처리하는 한계를 보이며, 나머지 프로세스에서는 데이터 처리를 하지 못하는 한계가 있다. 이러한 문제를 해결하기 위해 개선된 Python에서는 다중 프로세스로 처리 가능함에 따라 이를 이용한 채용 공고 수집기를 개발을 제안한다. 제안한 다중 프로세스를 사용한 수집기는 단일 프로세스보다 처리성능이 최대 3.42배 향상되었음을 확인하였다. 따라서, 다중 프로세스를 이용하여 채용 공고를 수집한다면 기존보다 더 빠른 데이터 처리와 데이터 수집 능력을 갖췄음을 확인하였다.

키워드: 크롤링(Crawling), 다중 프로세스(Multi Processing)

I. 서 론

최근, 취업포털 사이트의 방문 비중이 계속 증가하고 있다[1]. 제안한 채용공고 수집기는 구직을 희망하는 이용자들에게 신속한 정보를 제공하기 위해 대중에게 널리 알려진 Python 언어로 데이터를 수집한다.[2][3]. 하지만, 다중 프로세스 기반을 갖춘 컴퓨터가 보급됨에 따라, 다중 프로세스 기반을 갖춘 컴퓨터에서 동작하는 기존 Python에서는 단일 프로세스만을 지원하기 때문에 나머지 프로세스를 사용하지 못하는 문제가 있다.[4] 그러나, 개선된 Python 에서는 다중 프로세스를 지원하여 기존 Python 보다 더 빠른 데이터 처리 능력을 가진다[5]. 따라서 다중 프로세스를 이용하여 채용공고 정보를 수집한다면 기존보다 더 빠른 데이터 처리와 데이터 수집 능력을 가질 수 있다. 본 논문은 Python 언어 기반의 다중 프로세스를 이용한 채용공고 단어 수집 방법을 제안한다.

II. 설계 과정

제안한 채용정보 수집기에서 채용정보를 수집하고 단어를 처리하는 과정은 [그림 1]과 같다.

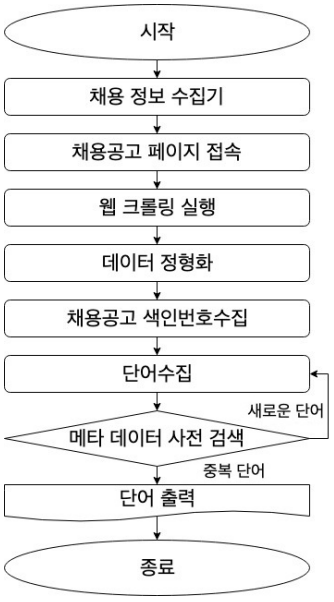

Fig. 1. 프로세스 흐름도

제안한 채용정보 수집기에서 실행하는 프로세스 개수는 최소 1개에서 최대 4개까지 지원가능하며, 개발환경은 아래 [표 1]과 같다. 수집 과정 중에 존재하는 메타데이터 사전은 별도로 용어사전에서 분류한다[6]. 그리고 자체적으로 수집한 채용공고 페이지에서 수집한 단어와 연관되는지 비교 후 단어 출력을 통해 결과를 나타내었다.

III. 실험

Table 1. 개발환경

CPU	Intel i7 2.6Ghz (4 Core/ 8 Thread)
RAM	16GB
개발언어	Python 3.6

별도의 용어사전에서 분류한 IT분야의 메타데이터 사전의 단어는 총 6314개이고[6], 채용공고 사이트의 트래픽 및 방어적인 대응을 고려하여, 자체적으로 수집한 페이지는 4페이지, 200개의 채용공고로 실험하였다. 아래 [표 2]의 실험결과는 프로세스 수에 따른 처리 시간이다. n은 프로세스의 수를 나타낸다.

Table 2. 실험 결과

	n = 1	n = 2	n = 3	n = 4
처리 시간 (초)	130	63	62	38

[표 2]에서 보는 바와 같이 프로세스 1개를 사용할 때보다 프로세스 4개를 사용할 때 3.42배 향상되었음을 확인할 수 있다.

VI. 결 론

본 논문에서는 Python 언어 기반의 다중 프로세스를 이용하여 채용공고 단어 수집방법을 제안하였다. 다중 프로세스로 처리하였을 때 기존보다 3.42배 더 나은 성능을 나타냄을 확인할 수 있었다. 향후 개발은 채용 공고를 분야별로 확대하여 키워드에 맞는 통합 채용 공고 수집기를 개발할 계획이다.

REFERENCES

[1] Nielsen Koreanclick, 취업포털 PC웹사이트 방문페이지별 비중, url:http://www.koreanclick.com/english/insights/newsletter_view.html?code=topic&id=503&page=1

[2] 채용공고 빅데이터 기반 SW분야 직업 및 직무변화 분석, 김정민(Jungmin Kim), 한국정보과학회, 2018

[3] PYPL PopularitY of Programming Language, url:http://pypl.github.io/PYPL.html

[4] Python 2.5 Release, url:https://www.python.org/download/releases/2.5/

[5] Python 2.6 Multiprocessing, url:https://docs.python.org/2.6/library/multiprocessing.html

[6] 한글형태소 사전 NIADic (정보·통신) 키워드 추출, K-ICT 빅데이터 센터, url:https://kbig.kr/portal/kbig/knowledge/files/bigdata_report.page?bltnNo=10000000016451