Journal of Knowledge Information Technology and Systems(JKITS), Vol. 15, No. 6, pp. 1075-1083, December 2020
https://doi.org/10.34163/jkits.2020.15.6.015

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700 (Print), ISSN 2734-0570 (Online)

http://www.kkits.or.kr

A Design and Implementation of Paragraph-based Focused Web Crawler Using Semantic Priority of Link

Nam-Oh Kang[1], **Jae-Ho Kim**[*2]

[1]*College of Humanities and International Studies, KeiMyung University*
[2]*Department of Computer Science and Engineering, Gangneung-Wonju National University*

ABSTRACT

A search engine maintaining whole Web consistency is very important to retrieve information correctly and efficiently. However, as the size of Web is rapidly growing and content is also dynamically changing, it is impossible for the search engine to achieve the goal by using limited resources such as hardware, network and computing time. In order to solve this problem, a focused web crawler has been introduced which can identify and visit the most promising links related to a specific topic and avoid downloading off-topic documents efficiently under limited resources. In this research, we propose a paragraph-based focused web crawler using semantic priority of link. The proposed system selects promising links from a downloaded web page by measuring similarity between a topic and link's data such as anchor text and a paragraph containing the link. In this paper, different from existing methods, we proposed a novel similarity function for calculating a link priority by using WordNet. And we introduced a method to visit high-priority link first. We conducted experiments to prove the performance of the proposed paragraph-based web focused crawler by using some topics. The experimental result showed the paragraph-based web focused crawler using semantic priority of link improves the term frequency of document retrieval.

KEYWORDS: Web search engines, Focused web crawlers, Link priorities, Semantic webs, Information retrievals, WordNet

ARTICLE INFO: Received 8 October 2020, Revised 9 November 2020, Accepted 11 December 2020.

[*]Corresponding author is with the Department of Computer Science and Engineering, Gangnenug-Wonju National University, 150 Namwon-ro Heungup-myon, Wonju, 26403, KOREA.
E-mail address: kimjaeho@gwnu.ac.kr

1. 서 론

1990년대 팀 버너스 리가 월드 와이드 웹(World Wide Web)을 제안한 이후 웹은 인터넷의 대표적인 정보 공유 서비스로 자리 잡았다. 현재 전 세계 웹 사이트 수는 18억 개를 넘어섰고[1], 이런 방대한 웹 공간에서 정보를 효과적으로 검색하기 위해 사람들은 검색 엔진(Search engine)을 이용한다.

검색 엔진은 사용자의 검색 질의에 빠른 응답을 위해 웹 크롤러(혹은 스파이더, 에이전트, 로봇)을 이용해서 미리 대량의 데이터를 수집하고 엄청난 양의 웹 페이지를 색인화 해두고 있다. 예를 들어서 구글의 경우 검색 색인은 수십억 개의 웹페이지를 포함하고 있으며 그 크기는 100페타바이트를 넘어서고 있다[2]. 그런데도 검색 엔진이 최신의 웹 상황을 반영하기에는 한계점이 있는데 그 이유는 웹을 통해 공개되는 웹 페이지의 수와 크기는 급속도로 증가하고 그 내용 또한 매우 동적으로 변하기 때문이다[3].

검색 엔진이 최신의 웹 상황을 유지하기 위해서는 웹 크롤러를 이용해서 자주 웹 크롤링을 해주어야 한다. 하지만 검색 엔진에서 사용하는 일반적인 웹 크롤러는 네트워크 대역폭, 시간, 그리고 저장소 등의 제약으로 인해 웹상의 모든 페이지를 자주 내려받을 수는 없다. 이런 문제를 극복하기 위해 1999년 S. Chakrabati는 집중 웹 크롤러(Focused web crawler)를 제안했다[4].

웹 공간 전반을 대상으로 검색된 정보의 색인을 구축하는 일반적인 웹 크롤러와 달리 집중 웹 크롤러는 웹에서 특정 영역의 주제를 검색하거나 사용자가 관심 있어 하는 정보만을 수집 및 관리한다. 따라서 집중 웹 크롤러는 일반 웹 크롤러보다 훨씬 적은 자원을 사용하며 실시간적인 검색 정보를 제공해 준다는 장점으로 인해 정보 추출, 디지털 도서관, 그리고 텍스트 분류 등의 응용에 활용된다[5].

집중 웹 크롤러는 획득한 문서를 분석, 주제와 관련성이 없는 링크(URL)는 배제하고, 관련성이 높은 링크는 추출 및 방문함으로써 효과적인 정보 검색 및 성능 향상을 목표로 한다[6,7]. 그리고 이 목표의 달성을 위해 웹 문서와 주제의 연관성 측정, 웹 문서로부터의 텍스트 및 링크 추출, 주제와 링크에 연결된 문서의 연관도 예측, 링크들의 방문 알고리즘 등 다양한 연구가 수행되었다.

본 논문에서는 집중 웹 크롤러의 성능 향상을 위해 웹 페이지를 링크 중심의 문단으로 추출하고, 사용자가 입력한 주제와 링크의 앵커 텍스트 및 문단과의 의미적 유사성을 측정, 방문할 링크의 우선 순위를 결정하는 의미 활용 문단 기반 집중 웹 크롤러를 소개한다. 실험을 통해 본 논문에서 제시하는 웹 크롤링 방식이 기존의 웹 크롤링 방식보다 나은 성능을 보였다.

본 논문의 구성은 다음과 같다. 제2장에서는 웹 크롤러 구축을 위해 사용된 기존의 기법들에 관해서 설명한다. 제3장에서는 링크 중심 웹 문서의 문단화, 링크의 중요도 측정 그리고 이를 이용한 의미 활용 문단 기반 집중 웹 크롤러 구축에 대해서 살펴본다. 그리고 제4장에서는 실험을 통해 제안한 시스템의 성능을 알아보고 마지막으로 제5장에서는 결론을 맺는다.

2. 관련 연구

2.1 집중 웹 크롤러

집중 웹 크롤러는 주어진 주제와 관련된 웹 페이지를 최대한 그리고 효과적으로 수집하는 것을 목표로 한다. 이를 달성하기 위해 집중 웹 크롤러는 이미 수집된 웹 페이지들의 우선순위를 정하고 이들로부터 방문할 링크들을 추출한 후 해당 링크

의 웹 페이지를 내려받기 전에 주제와의 관련성을 예측하여서 관련성이 높은 웹 페이지를 수집하고 관련성이 낮은 링크는 배제하도록 설계된다.

다음의 <그림 1>은 이전의 연구들[5-8]에서 제시된 집중 웹 크롤러의 일반적인 구조를 보여준다.

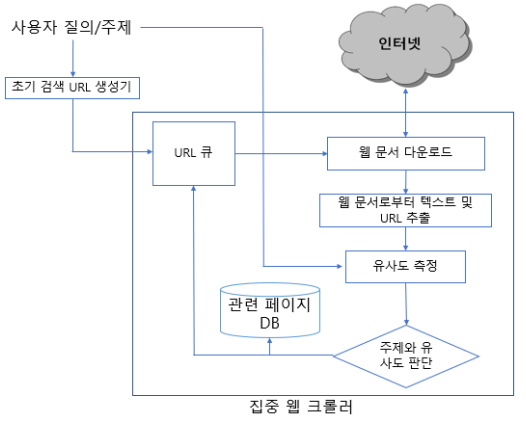

그림 1. 집중 웹 크롤러 구조
Figure 1. Structure of Focused Web Crawler

집중 웹 크롤러의 동작은 다음과 같다.

1. 사용자가 질의어 혹은 주제를 입력하면 초기 검색 URL 생성기(일반적으로 검색 엔진 활용)는 검색 URL을 생성하여 URL 큐에 초기 방문 URL(링크)을 입력한다.
2. 크롤러는 URL 큐로부터 링크를 가져와서 해당 웹 페이지를 다운 받는다.
3. 다운 받은 웹 페이지로부터 텍스트와 링크를 분리해 낸다.
4. 사용자가 입력한 질의어 혹은 주제와 텍스트 및 링크 앵커 텍스트 사이의 유사도 측정을 통해 링크의 중요도를 측정한다.
5. 크롤러는 주제와 관련성이 높게 측정된 링크들을 URL 큐에 입력하고 해당 페이지를 관련 페이지 DB에 저장한다.
6. 2-5를 반복 수행하며 URL 큐가 비거나 정해진 검색 결과 리스트가 만족되면 수행을 멈춘다.

기존의 많은 연구가 위에서 서술한 집중 웹 크롤러에서 특정 동작의 개선을 통해 집중 웹 크롤러의 성능 향상을 도모했다.

2.2 링크 중요도 측정

수집된 웹 문서에서 주제와 관련 있는 링크를 선별해내고 관리하는 것은 집중 웹 크롤러의 성능 향상에 중요한 역할을 한다.

[9-12]에서는 검색어와 웹 페이지 전체의 유사도 측정보다 웹 페이지를 콘텐츠 블록으로 나눈 후 블록 단위로 유사도를 측정하는 것이 더욱 좋은 결과를 산출함을 보여주었다. [5]에서는 이런 결과를 바탕으로 검색 주제와 연관도가 높은 콘텐츠 블록 내의 링크를 추출 및 방문에 활용하는 링크 중요도 측정(Link Priority Evaluation) 방법을 제안했다. 그리고 링크의 앵커 텍스트가 때로는 목적지 웹 페이지를 요약하지 않음으로 인해 발생할 수 있는 링크와 주제와의 유사도 측정 오류를 완화하기 위해, JFE 유사도 측정 방법을 제안했다.

$$Sim_{JFE}(u,v) = \lambda * Sim_{anchor}(u,v) + (1-\lambda) * Sim_{context}(u,v) \quad (1)$$

u: 링크

v: 주제

$Sim_{anchor}(u,v)$: 링크의 앵커 텍스트와 주제와의 유사도

$Sim_{context}(u,v)$: 링크를 포함한 콘텐츠 블록과 주제와의 유사도

λ: 중요도 인자 ([5]에서 기본적으로 0.5 설정)

3. 의미 활용 문단 기반 집중 웹 크롤러

본 논문에서 제안하는 링크의 의미 중요성을 이용하는 문단 기반 집중 웹 크롤러는 2.1절에서 언급했던 기존의 집중 웹 크롤러에 비해 웹 페이지의 문단 단위 텍스트 추출, 링크의 중요도 측정 및 주제와 문단 관련성 기반 링크 방문 등에서 차이가 있다.

제안하는 의미 활용 문단 기반 집중 크롤러는 다음과 같은 구조를 가진다.

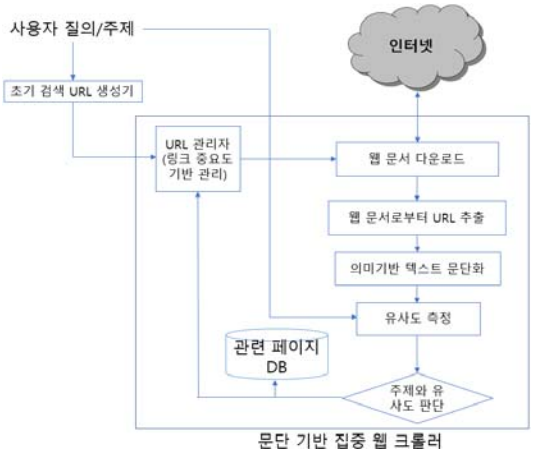

그림 2. 의미 활용 문단 기반 집중 웹 크롤러 구조
Figure 2. Structure of Paragraph-based Focused Web Clawer using Semantic Priority of Link

1. 사용자가 질의어 혹은 주제를 입력하면 초기 검색 URL 생성기(일반적으로 검색 엔진)는 검색 URL을 생성하여 URL 관리자에 초기 방문 링크(URL)을 입력한다. 초기 방문 URL의 중요도는 최대값을 가진다.
2. URL 관리자로부터 가장 중요도가 높은 링크를 가져와서 해당하는 웹 문서를 인터넷으로부터 다운로드 한다.
3. 다운 받은 웹 문서로부터 모든 링크들을 추출한다. 그리고 추출된 링크를 중심으로 웹 페이지의 문단화를 수행한다. 이에 대해서는 3.1절에서 설명한다.
4. 사용자가 입력한 질의어 혹은 주제와 문단화 된 텍스트 및 링크 앵커 텍스트 사이의 의미적 유사도 측정을 통해 링크의 중요도를 계산한다. 이에 대해서는 3.2절에서 설명한다.
5. 크롤러는 주제와 관련성이 높게 측정된 링크들을 URL 관리자에 입력하고 해당 페이지를 관련 페이지 DB에 저장한다. 이때 URL 관리자는 링크들을 중요도를 바탕으로 새롭게 정렬한다. 이에 대해서는 3.3절에서 설명한다.
6. 2-5를 반복하여 수행하며 관리하는 링크들이 더 이상 없거나 정해진 검색 결과 리스트가 만족되면 수행을 멈춘다.

3.1 링크 중심 문단화

웹 문서는 텍스트 내용, 링크, 광고 및 네비게이션 정보 등으로 이루어져 있다. 이런 웹 페이지로부터 비 본문 영역을 제거하고 텍스트 및 링크를 추출하기 위한 다양한 방법이 제시되었다.

[6]에서는 문서 내의 각각 텍스트 블록들이 본문 영역에 해당하는지 분류하기 위해 의사결정트리를 생성하고 이용했으며 분류를 위한 특징으로 텍스트 블록의 단어 및 링크 밀도와 HTML 태그 분포 및 텍스트 블록 간 거리 등을 포함하는 문맥 정보를 사용했다. 하지만 이는 텍스트와 링크를 각각 추출하며 문단화를 위한 별도의 작업을 거쳐야 한다는 단점이 있다.

본 연구에서는 획득한 웹 문서를 DOM으로 변환 후 XPath를 이용해서 링크와 링크를 포함한 문단을 추출한다. 이 과정에서 링크들 중 내부 문서 내비게이션을 위한 링크나 앵커 텍스트가 없는 링크들은 제거된다.

3.2 링크의 중요도 계산

3.1에서 추출된 링크들은 각각이 주제와 얼마나 의미적 연관성이 높은지에 대한 유사도를 계산하고 이를 링크의 중요도 값으로 활용한다.

본 연구에서는 웹 페이지로부터 주제와 연관성 있는 링크를 추출하기 위해 주제와 링크의 앵커 텍스트 및 링크를 포함하는 문단과의 의미적 유사도를 측정한다. 즉 링크의 앵커 텍스트 및 이를 포함하는 문단이 의미적으로 주제와 가까울수록 해당 링크를 중요한 링크로 판단한다. 주제의 유사도를 바탕으로 한 링크의 중요도를 계산하기 위해서는 단어 간의 유사도 측정이 가능해야 하는데 본 연구에서는 이를 측정하기 위해 WordNet[13]을 활용한다. 본 연구에서 사용한 주제와 링크 사이의 의미적 유사도 측정은 다음과 같다.

t: 주제, l: 링크, a: 앵커 텍스트, p: 문단 (2)
w: 단어, tw: 주제어의 단어
aw: 앵커 텍스트의 단어, s: 문장

$$t = \{tw_1, tw_2, \cdots\cdots, tw_n\}$$

$$a_l = \{aw_1, aw_2, \cdots\cdots, aw_n\}$$

$$p_l = \{s_1, s_2, \cdots\cdots, s_n\}$$

$$s_i = \{w_{i1}, w_{i2}, \cdots\cdots, w_{in}\} = w_i$$

$$Sim_{anchor}(t,l) = \frac{\sum_i \sum_k path\_similarity(tw_i, aw_k)}{|t|+|a|}$$

$$Sim_{paragraph}(t,l) = \frac{\sum_i \sum_j \sum_k path\_similarity(tw_i, w_{jk})}{|t|+|p|}$$

$Sim_{anchor}(t,l)$는 주제와 앵커 텍스트의 유사도를, $Sim_{paragraph}(t,l)$는 주제와 링크를 포함하는 문단의 유사도를 구한다. WordNet의 path_similarity 함수는 두 단어의 의미적 거리 유사도를 구한다. 그리고 주제와 링크의 유사도 Sim(t,l) 측정은 $Sim_{anchor}(t,l)$과 $Sim_{paragraph}(t,l)$의 조화 평균을 이용한다.

$$Sim(t,l) = \frac{1}{\frac{1}{Sim_{anchor}(t,l)} + \frac{1}{Sim_{paragraph}(t,l)}} \quad (3)$$

집중 웹 크롤러의 링크 중요도 계산 방식은 시스템의 성능에 큰 영향을 준다. 식 (1)의 JEF에서는 링크의 중요도를 계산할 때 λ를 이용하여 주제-앵커 텍스트 유사도와 주제-콘텐츠 블록 유사도 간에 비중을 조절한다. 하지만 [5]에서는 λ를 0.5로 정함으로 인해 서로 다른 기준에서 측정된 두 유사도를 단지 산술 평균하고 있으며, λ의 합리적 설정 방식 또한 제안하지 못하고 있다. 본 연구에서 제안하는 링크의 중요도 계산 방식은 주제-앵커 텍스트와 주제-콘텐츠 블록의 유사도를 조화 평균을 이용한다. 이는 링크의 앵커 텍스트나 링크를 포함하는 문단 중 하나라도 주제와 의미적 유사성이 떨어질 경우 링크의 중요도 값을 떨어뜨리게 한다. 즉 두 요소 모두 높은 유사도를 가질 때 링크의 중요도 또한 높은 값을 가지도록 동작한다.

그리고 본 연구에서 제안하는 링크의 중요도 계산은 WordNet의 path_similarity를 이용함으로써 유사도 측정에 단어 간의 의미 유사성을 반영하여 계산하도록 설계하였다. 이는 기존의 단어 빈도수를 기반한 유사도 계산 방식보다 단어의 의미성을 기반한 유사도 측정이 가능하게 한다.

3.3 링크 중요도 기반 웹 페이지 방문

3.2에서 구한 주제와 링크 간의 유사도는 링크

의 중요도로써 저장되고 집중 웹 크롤러가 방문할 웹 페이지를 선택하는데 이용된다. 웹 크롤러는 웹 페이지를 획득하고 웹 페이지 내의 링크들을 추출 후 중요도를 계산 후 중요도를 중심으로 방문할 링크들의 정렬을 실행한다. 이런 작업은 크롤러가 깊이 우선(Depth-first) 크롤링[14]이나 넓이 우선(Breath-first) 크롤링[14]이 아닌 주제와 링크의 유사도를 중심으로 한 크롤링이 가능하도록 해준다.

4. 실험

주제와 문단간의 의미 비교를 이용하는 문단 기반 집중 웹 크롤러의 성능을 검증하기 위해 실험을 수행했다. 실험에서는 [5]에서 사용했던 검색 주제 중 일부를 이용했으며 그 대상은 다음과 같다.

표 1. 주제와 시작 URL
Table 1. The seed URLs for topics

주제	Seed URL
Computer	https://en.wikipedia.org/wiki/Computer
Basketball	https://en.wikipedia.org/wiki/Basketball
Military	https://en.wikipedia.org/wiki/Military

본 논문에서 제안하는 시스템(SPFC)과 [5]에서 제안하는 Cosine 유사도를 활용한 문단 기반 집중 웹 크롤러(PFC) 그리고 넓이 우선 탐색을 수행하는 기존의 집중 웹 크롤러(FC) 간의 성능 비교를 실시했다. 시스템들은 각 주제에 대해 20, 30, 40, 50개의 웹 페이지를 검색하고 각 페이지에서 해당 주제에 대한 단어 빈도수를 측정했다.

<그림 3>은 주제어 Computer에 대해 시스템들이 웹 크롤링한 결과의 단어 빈도수를 구한 것이다.

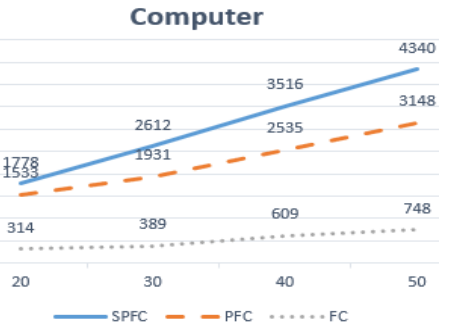

그림 3. 주제어 Computer에 대한 단어 빈도수 비교
Figure 3. Comparison of term frequency for Topic Computer

<그림 4>와 <그림 5는>은 주제어 Military와 Basketball에 대해 시스템들이 웹 크롤링한 결과의 단어 빈도수를 구한 것이다.

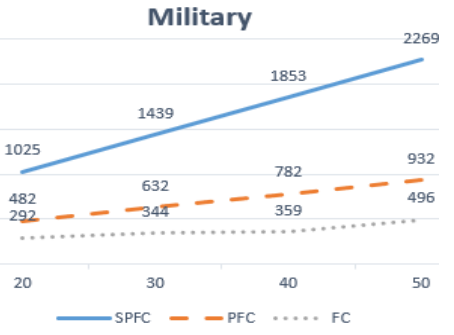

그림 4. 주제어 Military에 대한 단어 빈도수 비교
Figure 4. Comparison of term frequency for Topic Military

위의 실험들에서 보듯이 의미 활용 문단 기반 집중 웹 크롤러가 다른 시스템들에 비해 좋은 단어 빈도수 결과를 산출했다.

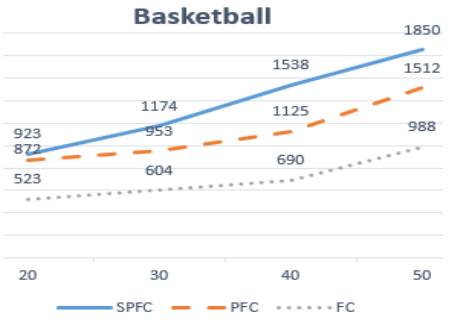

그림 5. 주제어 Basketball에 대한 단어 빈도수 비교
Figure 5. Comparison of term frequency for Topic Basketball

5. 결 론

집중 웹 크롤러의 성능을 향상하기 위해 의미 활용 문단 기반 집중 웹 크롤러를 설계 및 구현했다. 제안된 시스템은 획득한 웹 페이지로부터 주제와 관련성 있는 링크를 추출하고 링크와 주제와의 유사도를 중심으로 웹 페이지를 방문함으로써 주제와 관련성이 높은 페이지를 효과적으로 확보했다. 주제와 링크의 관련성을 계산하기 위해 WordNet을 이용, 주제와 링크의 데이터 간에 유사도를 측정했다. 측정된 유사도는 링크의 중요도로 활용되어 웹 크롤러가 방문할 우선순위를 정하는데 사용, 웹 크롤러의 성능을 향상 시켰다. 몇몇 주제를 집중 크롤링하는 것으로 실험을 하였고, 실험 결과에서는 의미 활용 문단 기반 집중 웹 크롤러가 단어 빈도수 측면에서 좋은 결과를 보였다.

본 시스템의 성능 향상을 위해 향후 다음 연구를 수행할 것이다. 첫째 특정 주제에 특화된 집중 크롤러 구축을 위해 온톨로지의 도입[15-17]을 시도한다. 온톨로지는 특정 영역의 개념이나 지식을 표현하기 위해 사용된다. 본 연구에서 문단이나 주제의 의미적 유사성 파악을 위해 WordNet을 활용했다. 하지만, 집중 웹 크롤러가 특정 분야의 온톨로지를 활용한다면 보다 나은 검색 결과를 산출할 것이다. 둘째 링크의 중요도 판단에 기계 학습의 도입을 시도한다. 집중 웹 크롤러는 시작 URL부터 다양한 웹 페이지를 획득하고 분석하며 링크의 중요도를 계산한다. 이러한 과정을 통해 문단에 사용된 단어의 패턴을 학습하고 활용한다면 보다 효율적인 집중 웹 크롤러의 구축이 가능할 것이다.

References

[1] Total number of websites, https://www.internetlivestats.com/total-number-of-websites/, Aug. 2020.

[2] How search organizes information, https://www.google.com/intl/en/search/howsearchworks/crawling-indexing/, Aug. 2020.

[3] S. W. Kim, *Focused crawler for efficient web gathering*, Soongsil University, 2007.

[4] S. Chakrabarti, M. van den Berg, and B. Dom, *Focused crawling: a new approach to topic-specific web resource discovery*, In proceedings of 8th International World Wide Web Conference, pp. 545-562, 1999.

[5] H. Lu, D. Zhan, L. Zhou, and D. He, *An improved focused crawler: using web page classification and link priority evaluation*, Mathematical Problems in Engineering, Vol. 2016, pp. 1-10, 2016.

[6] S. Shah, S. Patel, and S. Nair, *Focused and deep web crawling-A review*, International Journal of Computer Science and Information Technologies, Vol. 5, No. 6, pp. 7488-7492, 2014.

[7] D. Bhatt, D. A. Vyas, and S. Pandya, *Focused web crawler*, Advanced in Computer Science and Information Technology, Vol. 2, No. 11, pp. 1-6, Apr.-Jun. 2015.

[8] N. W. Min and A. N. Hlaing, *Ranking hyperlinks approach for focused web crawler*,

International Conference on Advances in Engineering and Technology, pp. 233-235, Mar, 2014.

[9] T. Peng and L. Liu, *Focused crawling enhanced by CBP-SLC*, Knowledge-Based Systems, Vol. 51, pp. 15-26, 2013.

[10] N. Luo, W. Zuo, F. Yuan, and C. Zhang, *A new method for focused crawler cross tunnel*, First International Conference, RSKT, Vol. 4062 of Lecture Notes in Computer Science, pp. 632-637, Springer, 2006.

[11] T. Peng, C. Zhang, and W. Zuo, *Tunneling enhanced by web page content block partition for focused crawling*, Concurrency Computation Practice and Experience, Vol. 20, No. 1, pp. 61-74, 2008.

[12] S. J. Park, and J. H. Kim, *Paragraph-based k-means clustering by using meaning-based paragraph division*, Journal of Knowledge Information Technology and Systems, Vol. 12, No. 1, pp. 157-164, Feb. 2017.

[13] WordNet, https://wordnet.princeton.edu/, Sep. 2020.

[14] B. Ganguly, and R. Sheikh, *A review of focused web crawling strategies*, International Journal of Advanced Computer Research, Vol. 2, No. 4, pp 261-267, Dec. 2012.

[15] Ontology, https://www.w3.org/standards/semanticweb/ontology, Sep. 2020.

[16] G. Subbiah, M. Jayaraj, V. Kalyan, SrinivasaMurthy, and G. Aghila, *Ontology-based web crawler*, Proc. ITCC: Coding Comupt., pp. 334-341, 2004.

[17] J. Lee, *The ontology construction of Instructional domain knowledge*, Journal of Knowledge Information Technology and Systems, Vol.11, No. 1, pp. 57-63, Feb. 2016.

링크의 의미 중요성을 이용하는 문단 기반 집중 웹 크롤러 설계 및 구현

강남오[1], 김재호[2]

[1]*계명대학교 인문국제학대학 강사*

[2]*강릉원주대 과학기술대학 컴퓨터공학과 교수*

요 약

검색 엔진과 전체 웹의 일치성 유지는 정보를 정확하고 효과적으로 검색하는데 아주 중요하다. 하지만 웹의 크기가 빠르게 성장하고 그 내용이 또한 동적으로 변하기 때문에, 검색 엔진이 하드웨어, 네트워크 그리고 컴퓨팅 시간과 같은 제한된 자원을 이용해서 그와 같은 목표를 달성하는 것은 불가능하다. 이런 문제를 해결하기 위해, 제한된 자원하에서도 효율적으로 주제와 관련성이 높은 링크를 발견 및 방문하고 관련성이 없는 문서의 검색을 회피하는 집중 웹 크롤러가 소개되었다. 본 연구에서 우리는 링크의 의미적 중요도를 이용하는 문단 기반 집중 웹 크롤러를 제안한다. 제안된 시스템은 주제와 링크의 앵커 텍스트 및 링크를 포함하는 문단과 같은 링크 데이터 간의 유사도 측정을 이용해서 획득한 웹 페이지로부터 가능성 있는 링크들을 선별한다. 본 연구는 다음의 점이 기존의 연구 방법과 차이가 있다. 우리는 링크 중요도 계산을 위해 WordNetd을 사용하는 새로운 유사도 함수를 제안했다. 그리고 높은 중요도를 가지는 링크를 우선 방문하는 방법을 제시했다. 제안된 시스템의 성능을 증명하기 위해 몇몇 주제와 시작 URL을 이용해서 실험을 수행했다. 실험의 결과는 본 논문에서 제안하는 링크의 의미 중요도를 사용하는 문단 기반 웹 집중 크롤러가 검색에 있어서 단어 빈도수의 향상을 보였다.

감사의 글

이 논문은 2019년도 강릉원주대학교 학술연구조성비 지원에 의하여 수행되었음.

Nam Oh Kang received the bachelor's degree, the M.S. degree and the Ph.D degree in the Department of Computer Science and Engineering from Chung-Ang University in 1997, 2002, and 2006, respectively. His current research interests include web crawler, artificial intelligence, and machine learning.

E-mail address: namohkang@gmail.com

Jae Ho Kim received the bachelor's degree, the M.S. degree and the Ph.D degree in the Department of Computer Science and Engineering from Chung-Ang University in 1988, 1990, and 2004, respectively. He has been a professor in the Department of Computer Science and Engineering at Gangneung-Wonju National University since 1997. His current research interests include web crawler, artificial intelligence, and machine learning, He is a life member of the KKITS.

E-mail address: kimjaeho@gwnu.ac.kr