신경망을 적용한 온톨로지 기반의 Focused Crawling

ZhengHai-tao 강보영 남궁현 김홍기
서울대학교 지식공학연구실
{quickly, comeng99, ngh, hgkim}@snu.ac.kr

Ontology-Based Focused Crawling Combined with Neural Network

Hai-tao Zheng Bo-young Kang Hyun Namgoong Hong-Gee Kim
Biomedical Knowledge Engineering Lab, Seoul National University

요 약

Focused crawling은 검색시스템의 구축을 위한 웹 문서 수집단계에서, 미리 정의된 토픽 집합들과 관련성을 가지는 웹 문서를 수집하기 위하여 제안되었다. 이러한 focused crawling 연구에서 보다 효과적인 웹 문서 수집을 위해 주어진 토픽에 대한 양질의 배경지식을 제공할 수 있도록 온톨로지가 활발히 활용되어왔다. 그러나 기존의 온톨로지 기반 focused crawling 연구는 토픽과 웹 문서 간의 관련성 측정을 위하여, 주어진 토픽과 관련있는 온톨로지 내 각 개념들에 직관에 의존한 가중치를 부여하여 활용하였다. 하지만 이러한 직관에 의존한 가중치부여 기법은 안정된 수집결과를 도출할 수 있는 최적화된 가중치 값을 얻기가 힘든 한계가 있다. 따라서 본 논문에서는 이러한 개념에 대한 가중치가 학습에 의하여 자동으로 결정되도록, 인공신경망을 적용한 온톨로지 기반 focused crawling 기법을 제안한다. 웹 상에서 제안된 시스템의 성능을 실험한 결과 기존의 온톨로지 기반 수집 기법에 비하여 보다 향상된 결과를 보임을 알 수 있었다.

1. 소개

인터넷의 발달과 웹 정보의 비약적인 증가로 인하여 유용한 정보의 정확한 검색은 갈수록 힘들어 지고 있으며, 이러한 관점에서 focused crawling과 같은 지능화된 문서 수집 기법의 개발이 중요한 의미를 가지게 되었다[3,6,7]. 전통적으로, 문서 수집 기법은 데이터 스토리지 구축을 위한 기본적인 기술로서 발전되어 왔으며, 최근에는 미리 정의된 탐색 토픽(topic) 집합과 관련 있는 웹 문서의 탐색을 위해 이용되고 있다.

Focused crawling 연구의 주된 이슈는 문서의 수집과정에서 어떻게 효율적으로 주어진 토픽과 관련된 분야를 탐색하고 얼마나 높은 수확율(harvest rate)을 가지도록 하는가이다. 이러한 관점에서 보다 효과적인 웹 문서 수집을 위하여 주어진 토픽에 대한 양질의 배경지

[1] 본 연구는 보건복지부 온톨로지 기반의 EHR 상호운용 기술개발 과제(과제번호: A05 -0909-A80405-06A2-15010A)의 지원에 의해 이루어진 것임.

식을 문서 수집 과정에서 이용할 수 있도록, 특정 토픽에 대한 배경지식을 개념(concept)과 관계(relation)의 형태로 제공할 수 있는 온톨로지를 활용한 기법들이 활발히 제안되었다[6][8][10].

다양한 온톨로지 기반 문서 수집 기법 중, M.Ehrig[6]의 연구가 많은 연구의 초점이 되어왔다. 그러나 M.Ehrig의 방법은 토픽과 관련된 온톨로지 내 개념들에 사용자 직관에 의해 부여된 가중치 값을 활용하여 문서와 탐색 토픽 간의 관련성을 측정함으로써, 안정된 수집결과를 유도할 수 있는 최적화된 가중치 값을 얻기가 힘든 한계가 있다

따라서 본 논문에서는 이러한 개념에 대한 가중치 부여가 학습에 의하여 자동으로 결정되도록, 인공신경망을 적용한 온톨로지 기반 focused crawling 기법을 제안한다. 논문에서 제시되는 방법은 기존의 방법과 유사하게, 토픽과 관련있는 온톨로지 내 개념 및 개념이 가지는 가중치 값을 문서 수집을 위한 배경지식으로 이용한다. 인공신경망은 학습데이터의 속성을 입력으로 신경망의 뉴런(Neuron)들이 가지는 가중치를 학습시켜 학습을 수행하는 알고리즘이다[9]. 논문에서 제안된 방법에서는 각 개념들이 해당 토픽으로 분류되기 위해 기여하는 정도를 나타내는 가중치는 신경망에 의해 학습되므로, 사용자의 직관이 아닌 학습 데이터로부터 최적화된, 객관적인 가중치 값을 유도할 수 있다는 점에서 유용할 것으로 보인다.

논문은 다음과 같은 순서로 서술된다. 2장에서는 기존의 문서 수집 방법들에 대해 설명하고, 3장에서 제안된 기법과 그 구조에 대해 자세히 설명한다. 4장에서는 제안된 기법 및 기존 기법들을 실험한 결과가 제시되며, 마지막으로 5장에서 본 연구를 결론 맺는다.

2. 관련연구

미리 정의된 탐색 토픽(topic) 집합과 관련 있는 웹 문서를 탐색하는 기법인 focused crawling은 다양한 기법을 적용하여 활발히 연구되어왔다. S.Chakrabarti은 유전자 알고리즘을 이용한 수집 방법을 개발하였으며[3], M.Diligenti는 웹 상의 문서들 간에 가지는 문맥정보 (Context)를 모델링하는 알고리즘을 이용한 방법을 제안하였다[5]. 이 방법에서는 그래프 형태의 모델을 이용하여 중요한 웹 문서들 사이에 나타나는 일반적인 링크구조를 모델링 할 수 있으며, 이를 이용하여 특정 문서와 해당 토픽간의 거리를 측정할 수 있다[7]. 또한 J.Rennie는 기계학습을 기반으로 하는 focused crawling기법을 제안하여, Q-학습알고리즘을 통해 관련성이 높은 웹 문서의 탐색을 가능하게 하였다[11]. 그러나 이러한 방법들은 온톨로지를 토픽을 위한 사전지식으로 활용하지 않은 기법들이다.

온톨로지는 개념과 개념들간의 관계를 포함하는 특정 분야에 대한 기술(description)의 집합으로서, 특정 토픽에 관련된 기술 정보를 획득하기에 용이한 지식베이스이다. 이러한 온톨로지를 특정 토픽에 관련된 문서수집을 위한 배경지식으로 제공하고자 하는 온톨로지 기반 문서 수집 기법이 [6][8][10]등에서 활발히 연구되었다.

이러한 온톨로지기반 기반 문서 수집 기법 중, M.Ehrig[6]의 연구가 많은 연구의 초점이 되어왔는데, 해당 연구에서 제안된 문서 수집은 다음의 3단계로 진행된다. (1) 개체 참조 구축: 주어진 토픽에 관련된 온톨로지 개념이 웹 문서에서 등장하는 빈도수를 계산한다. (2) 관련 정보 편집: 이 단계에서는 토픽에 관련된 온톨로지 내 개념들이 선택되며, 각 개념들이 탐색 토픽에 관련되는 정도를 나타내는 가중치 값이 정해진다. 이 가중치 값은 개념과 토픽과의 온톨로지 내에서의 거리(distance)와 사용자의 직관에 의한 감쇠 요소(discounting factor)가 적용된 수식(=감쇠요소^{거리})에 의해 결정된다. (3) 요약: 마지막 단계는 앞의 단계에서 얻어진 각 개념의 가중치를 이용하여, 웹 문서에서 발생한 개념의 빈도수 × 개념 가중치 값의 합을 웹 문서와 토픽간의 관련성 정도로 계산한다.

그러나 Ehrig이 제시한 이러한 문서 수집 방법에서는 각 개념의 가중치 값을 결정하는 과정에서 사용자 직관에 의존함으로써, 일관된 수집결과를 도출할 수 있는 최적화된 가중치 값을 얻기가 힘들다는 문제가 있다. 따라서 본 논문에서는 이러한 개념에 대한 가중치 부여

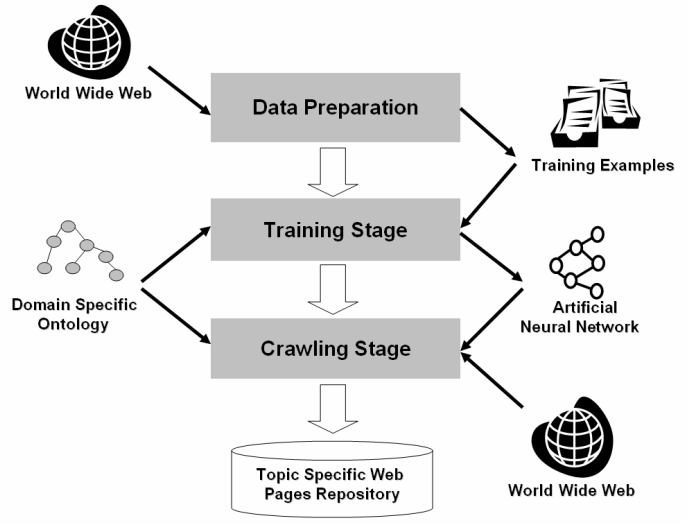
가 학습에 의하여 자동으로 결정되도록, 인공신경망을 적용한 온톨로지 기반 focused crawling 기법을 제안하며 보다 자세한 내용은 다음 장에서 다루도록 한다.

3. 신경망을 적용한 온톨로지 기반 Focused crawling

3.1 제안된 시스템 전체 구조

그림 1은 본 논문에서 제안한 시스템의 전체 구조를 보여준다. 그림 1에서 제안된 시스템은 데이터 준비, 학습, 문서 수집의 세 단계로 구성됨을 알 수 있다. 첫 번째 데이터 준비 단계는 주어진 토픽에 관련된 학습 문서를 구성하는 과정이다. 이 단계에서 사람에 의해 수동으로 판별된 각 문서와 해당 토픽에 대한 관련성이 학습을 위한 데이터로서 이용된다.

두 번째 학습 단계는 토픽에 관련된 온톨로지 내 개념들을 데이터 준비 단계에서 얻어진 각 학습 문서의



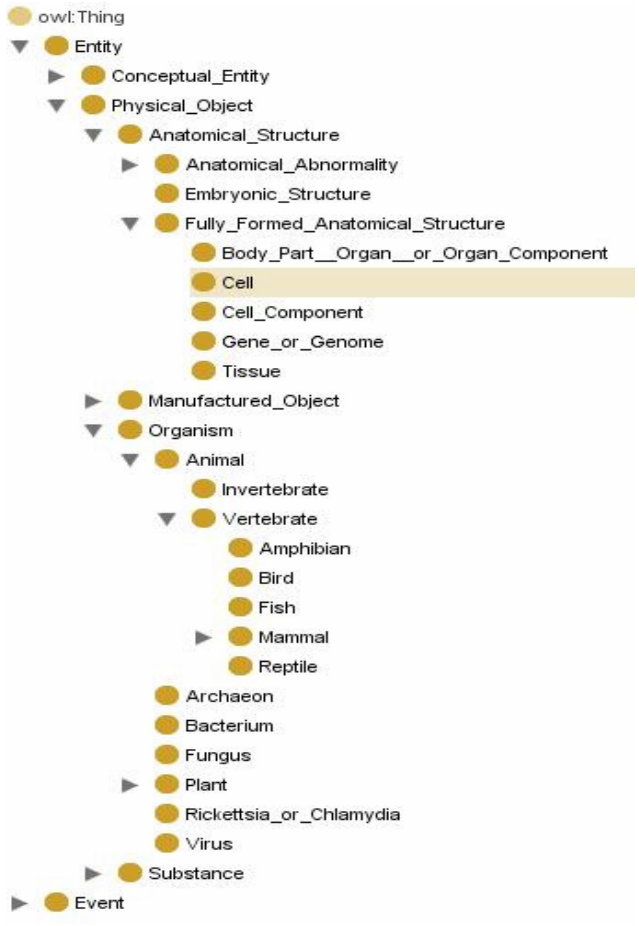
[그림 1] 제안된 시스템 전체 구조

학습 자질(feature)로 구성하여 시스템을 학습 시키는 작업을 수행하며 보다 자세히 설명하면 다음과 같다: 먼저, 온톨로지 내에서 주어진 토픽과 관련된 개념들이 선택된다. 이러한 개념들이 각각의 학습 데이터 문서들에서 가지는 출현 빈도수를 계산한 후 신경망을 위한 입력 값으로 학습한다. 이때 해당 문서가 주어진 토픽과 관련이 있으면 1, 그렇지 않으면 -1의 출력 값을 가진다. 이러한 방법으로 각 개념들이 해당 토픽으로 분류되기 위해 기여하는 정도를 나타내는 가중치 값을 신경망 학습에 의해 최적화된 객관적인 값으로 유도할 수 있다.

마지막으로 문서 수집 단계에서는 새로운 문서가 주어지면 학습된 신경망을 이용하여 문서를 수집할 것인지 판단한다. 두 번째 단계인 학습 단계에서와 마찬가지로 각 개념들이 주어진 문서에서 발생하는 출현 빈도수를 계산한 후 신경망을 위한 입력 값으로 입력하면, 학습된 지식에 기반하여 주어진 토픽에 대한 분류작업을 수행한다.

3.2 문서 수집 단계에서의 관련성 계산

본 절에서는 문서수집 단계에서 발생하는 웹 문서와 토픽 간의 관련성 계산 작업에 대하여 보다 자세한 소개를 다룬다. 먼저 관련성 계산을 위한 온톨지가 존재할 때, 주어진 토픽에 관련된 온톨로지 내 개념들은 다음과 같은 거리 개념에 의해 선택된다.

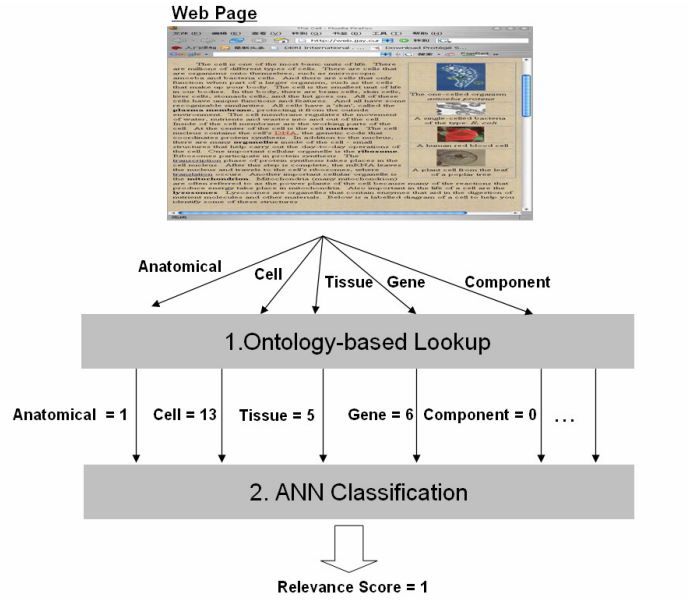


[그림 2] UMLS 온톨로지

정의1. (토픽과 개념 간의 거리) 토픽과 개념간의 거리는 $d(t, ci)=k$ 으로 표현될 수 있다. 여기에서 t 는 온토로지에 포함된 탐색 토픽, ci 는 온톨로지내의 개념을 의미하고, k 는 토픽과 개념간의 is-a 관계의 링크 수를 나타낸다.

토픽과 개념 간의 거리는 온톨로지 내에서 개념과 탐색 토픽간의 관련성을 의미하므로, 상대적으로 작은 $d(t, ci)$ 값을 가지는 개념 ci 은 탐색 토픽 t 와 높은 관련성을 가진다고 생각할 수 있다. 본 연구에서는 효율성과 수행시간을 고려하여 $d(t, ci) \leq 3$ 인 개념들만을 배경지식으로서 사용한다.

예를 들어, 본 연구의 실험에서 활용하는 의생명 온톨로지인 그림 2의 UMLS(Unified Medical Language System)[2] 온톨로지를 살펴보자. 주어진 토픽이 그림의 예에서는 Cell 이라고 가정하였을 경우, 온톨로지에서 탐색 토픽과 $d(Cell, ci) \leq 3$인 개념들을 선택하면, cell, component, gene, genome 등을 포함한 38개의 개념들이 선택된다. 이때, UMLS 온톨로지에는 Gene or



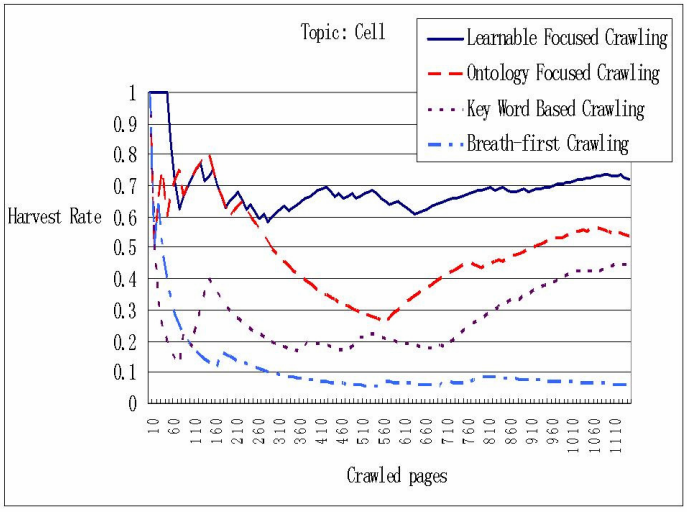
[그림 3] 관련성 계산 예제

Genome과 같은 복합어로 이루어진 개념들이 다수 존재하는데, 해당 용어들을 Gene와 Genome의 단일어로 분리하여 이용한다.

주어진 토픽에 대한 온톨로지 내 관련 개념들이 위와 같이 선택되면, 새로운 문서가 수집 시스템에 도착하였을 때 그림 3의 절차에 의해 토픽과 문서와의 관련성 정도가 계산된다. 즉, 각 개념들이 주어진 문서에서 발생하는 출현 빈도수를 계산한 후 신경망을 위한 입력값으로 입력하면 학습된 지식에 기반하여 주어진 토픽에 대한 분류작업이 수행된다. 예를 들어, 그림 3에서 주어진 토픽 Cell과 관련된 개념들이 anatomical, cell, tissue, gene, component 등으로 선택되었을 때, 주어진 문서에서 발생하는 출현빈도가 각각 1, 13, 5, 6, 0으로 계산되었다. 해당 개념 및 빈도수는 신경망의 입력으로 주어지고 그 결과 입력된 새로운 문서는 토픽 Cell과 관련 있는 문서로 분류된다.

4. 실험

본 장에서는 제안된 시스템의 성능 검증을 위하여 기존의 문서수집 방법들과의 실험 결과를 비교 분석하였다. 비교 대상이 되는 기법은 넓이 우선 탐색기반의 수집기(breadth-first search crawling)와 키워드 기반 수집기(focused crawling with simple keyword spotting), Ehrig에 의해 개발된 온톨로지 기반의



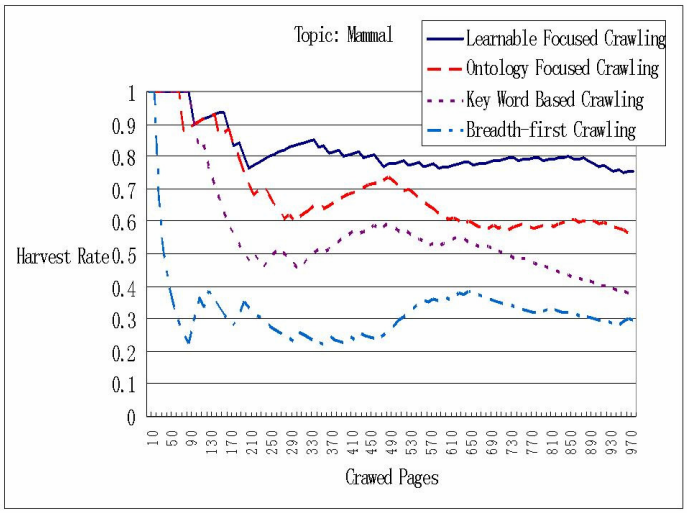
[그림 4] Cell을 탐색 토픽으로 한 수확율 그래프

focused crawling 기법이다[6]. 실험은 Cell 과 Mammal 토픽에 대해 수행하였다. 각 토픽에 대한 학습 데이터는 각각 약 100건의 웹 문서로 구성되어있으며, 학습데이터 구성을 위한 시작 웹 문서 주소로는

www.cellsalive.com, www.enchantedlearning.com/subjects/mammals 가 각각 사용되었다. 배경지식을 제공할 온톨로지로는 UMLS 온톨로지를 이용하였으며, 주어진 토픽과 개념들과의 거리인 $d(Cell, ci) \leq 3$ 으로 설정하였다. 웹문서 수집 성능 평가를 위한 척도로는 아래의 수확율(harvest rate)[1][3] 수식을 이용하였다. 수확율은 수집된 전체 웹문서 수(#p)에 대한 토픽에 관련있는 웹문서수(#p)의 비율을 나타내며, 수확율이 높을수록 문서의 웹 수집이 효과적으로 이루어 졌다고 평가할 수 있다.

$$hr = \frac{\sharp r}{\sharp p}, hr \in [0,1] \quad (1)$$

첫 번째 탐색 토픽인 Cell을 위하여 UMLS내에서 $d(Cell, ci) \leq 3$ 인 개념들을 추출한 결과 tissue, gene등을 포함한 38개의 개념들을 선택할 수 있었다. 토픽 Cell에 대한 웹 문서 수집 평가를 위한 시작 웹 문서의 주소는 http://web.jjay.cuny.edu/acarpi/NSC/13-cells.htm을 사용하였다. 토픽 Cell에 대한 수확율 실험결과는 그림 4와 같다. 그림 4에서 일련의 입력 웹 문서에 대하여 제안된 시스템이 기존 기법보다 보다 높은 수확율 성능을 보임을 알 수 있다.



[그림 5] Mammal을 탐색 토픽으로 한 수확율 그래프

두 번째 실험에서는 탐색 토픽인 Mammal을 위해서 UMLS내에서 $d(Mammal, ci) \leq 3$ 인 56개의 개념들을 추출 하였으며, 웹 문서 수집을 위한 시작 웹 문서의 주소는 http://www.ucmp.berkeley.edu/mammal/mammal.html 이다. 탐색 토픽 Mammal에 대한 실험 결과는 그림 5와 같으며, 토픽 Cell에 대한 그림 4의 결과와 유사하게 제안된 시스템이 기존 시스템에 비하여 좋은 성능을 보임을 알 수 있다.

방법 / 토픽	"Cell" 토픽	"Mammal" 토픽
Learnable focused crawling	0.6952	0.8215
Ontology focused crawling	0.4908	0.6892
Key word based crawling	0.3715	0.5641
Breadth-first crawling	0.1132	0.3252

[표 1] 방법별 평균수확율 비교

표 1은 탐색 토픽 Cell과 Mammal에 대한 방법별 평균 수확율을 보여준다. 평균수확율은 수집된 페이지의 수에 따른 수확율의 평균을 의미한다. Cell을 토픽으로 하였을 때 논문에서 제안된 시스템의 수확율은 기존의 온톨로지 기반의 문서 방법에 비해 평균적으로 0.2044만큼 더 정확한 결과를 보였으며 이는 키워드 기반의 방법보다는 0.3241, 넓이 우선 방법보다는 0.5820만큼 큰 수치이다. Mammal에 대한 수확율은 온톨로지 기반의 문서 방법에 비해서는 약 0.1323, 키워드 기반의 방법보다는 0.2574, 넓이 기반의 방법 보다는 0.4963만큼 정확한 결과를 보였다.

Cell과 Mammal 토픽에 대한 실험 결과로부터, 제안된 기법이 기존의 온톨로지 기반의 문서 수집 방법과 비교할 보다 나은 실험 결과를 보여줌을 알 수 있었다. 기존의 방법이 각 개념에 대해 사용자에 의해 주관적으로 부여된 가중치를 사용한다는 것을 고려할 때, 제안된 시스템에서는 학습에 의해 최적화된 가중치를 활용함으로써 기존 방법에 비해 보다 높은 성능을 보인 것으로 분석된다.

5. 결론

본 논문에서는 신경망을 적용한 온톨로지 기반의 focused crawling 방법을 제안하였다. 실험의 결과는 기존의 온톨로지 기반 방법에 비해 향상된 문서 수집 결과를 보여주었으며, 이러한 결과로부터 주어진 토픽

에 대하여 학습을 활용한 온톨로지 기반 웹 수집 기법이 기존 온톨로지 기반 기법 보다 효율적으로 문서를 수집함을 알 수 있었다. 향후 연구로서 제안된 기법을 보다 광범위한 토픽에 확대 적용하여 다양한 환경에서 일관되게 높은 성능을 가질 수 있도록 시스템을 개발하고자 한다. 또한 제안된 방법을 온톨로지를 사용하지 않고 단지 학습에 의존한 focused crawling 기법들과 비교 분석함으로써, focused crawling 분야에서 온톨로지의 활용 가능성에 대해 보다 심화하여 연구를 수행하고자 계획하고 있다.

참조문헌

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. *In WWW '10: Proceedings of the 10th international conference on World Wide Web*, pages 96-105, New York, NY, USA, 2001. ACM Press.

[2] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucl. Acids Res.*, 32(suppl 1):D267-270, 2004.

[3] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11-16):1623-1640, 1999.

[4] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[5] M. Diligenti, F. Coetzee, S. Lawrence, C. LeeGiles, and M. Gori. Focused crawling using context graphs. *In 26th International Conference on Very LargeDatabases*, Cairo, Egypt, pages 527-534, 2000.

[6] M. Ehrig and A. Maedche. Ontology-focused crawling of web documents. *In SAC '03: Proceedings of the 2003 ACM symposium on Applied computing, pages 1174*-1178, New York, NY, USA, 2003. ACM Press.

[7] C.-C. Hsu and F. Wu. Topic-speci¯c crawling on the web with the measurements of the relevancy context graph. *Inf. Syst.*, 31(4):232-246, 2006.

[8] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, and R. Volz. Ontology-focused crawling of documents and relational metadata. *In Proceedings of the Eleventh International World Wide Web Conference WWW-2002*, Hawaii, 2002.

[9] T. Mitchell. Machine Learning. *McGraw-Hill Science Engineering*, 1997.

[10] C. Su, Y. Gao, J. Yang, and B. Luo. An efficient adaptive focused crawler based on ontology learning. *In Hybrid Intelligent Systems, 2005. HIS '05. Fifth International Conference*, 6-9 Nov. 2005.

[11] J. Rennie and A. K. McCallum. Using reinforcement learning to spider the Web e±ciently. In I. Bratko and S. Dzeroski, editors, Proceedings of ICML-99, 16th International Conference on Machine Learning, pages 335-343, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.