J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

LOD 클라우드에서의 연결정책 기반 동일개체 심층검색 및 정제 시스템 구현

Implementation of Policy based In-depth Searching for Identical Entities and Cleansing System in LOD Cloud

김 광 민[1] 손 용 락[2*]
Kwangmin Kim Yonglak Sohn

요 약

본 연구에서는 동일연결트리플들을 생성하는 대신 각 LOD마다 연결정책을 수립, 공개하고 검색 시점에서 참조하는 방식으로 개체간의 동일성을 파악하는 방안과 이러한 연결정책을 명세하기 위한 어휘를 제안하였다. 또한, 연결정책이 운영되는 환경에서 여러 LOD들에 걸친 심층검색이 실질적으로 진행되는 것을 확인하기 위하여 PISC(Policy based In-depth Searching and Cleansing)을 구현하였으며 이를 Github에 공개하였다. LOD 클라우드는 여러 LOD들의 자발적인 참여로 이루어짐에 따라 검색된 개체들의 동일성에 대한 평가가 필요하다. 이에, PISC는 개체간 동일성 평가를 통하여 사용자가 요구한 동일수준 이상의 개체들로 정제된 검색결과를 제공한다. 검색결과로는 RDF로 모델링된 개체별 상세 검색내용과 이에 대한 의미적 구조인 온톨로지를 함께 제공된다. PISC에 대한 실험은 DBpedia의 5개 LOD를 대상으로 진행하였으며 소스와 타겟 RDF 트리플 목적어의 유사도를 0.9 정도로 요구할 경우 검색결과가 적절한 확장률과 포함률을 가지는 것으로 확인하였다. 또한, 연결정책에는 3개 이상의 타겟LOD를 명세할 경우 동일성이 충분히 검증된 개체들을 확보할 수 있는 것으로 확인하였다.

☞ 주제어 : 연결공개데이터, 동일개체검색, 연결정책, 온톨로지, 시멘틱웹

ABSTRACT

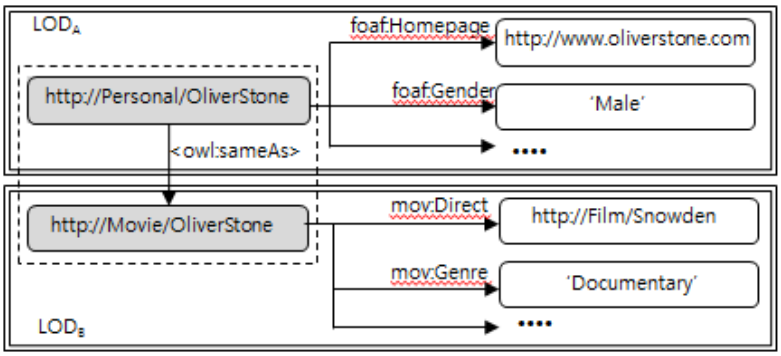
This paper suggests that LOD establishes its own link policy and publishes it to LOD cloud to provide identity among entities in different LODs. For specifying the link policy, we proposed vocabulary set founded on RDF model as well. We implemented Policy based In-depth Searching and Cleansing(PISC for short) system that proceeds in-depth searching across LODs by referencing the link policies. PISC has been published on Github. LODs have participated voluntarily to LOD cloud so that degree of the entity identity needs to be evaluated. PISC, therefore, evaluates the identities and cleanses the searched entities to confine them to that exceed user's criterion of entity identity level. As for searching results, PISC provides entity's detailed contents which have been collected from diverse LODs and ontology customized to the content. Simulation of PISC has been performed on DBpedia's 5 LODs. We found that similarity of 0.9 of source and target RDF triples' objects provided appropriate expansion ratio and inclusion ratio of searching result. For sufficient identity of searched entities, 3 or more target LODs are required to be specified in link policy.

☞ keyword : Linked Open Data, Searching Identical Entity, Link Policy, Ontology, Semantic Web

1. 서 론

오늘날의 웹은 문서의 웹 구조로 발전하였다. 그 결과 연결은 페이지단위로 이루어지며 페이지 내용에 대한 의미적 구조가 부재하여 상세 데이터단위에서의 활용에 많은 어려움이 있다[1]. 이를 극복하기 위하여 팀 버너스 리는 데이터의 웹인 시멘틱 웹 개념을 주창하였고[2] 이에 대한 구체적인 구현으로 2007년부터 LOD(Linked Open Data) 클라우드가 구축되기 시작하여 2018년 3월 기준으로 1,163개 LOD들이 참여하고 있다[3]. LOD에서는 존재하는 사실을 {주어, 술어, 목적어} 구조를 가지는 RDF(Resource Description Framework) 트리플 형식으로 실체화한다[4]. 즉, (그림 1)과 같이 LOD에서는 {<OliverStone> <Homepage> <www.oliverstone.com>}, {<OliverStone> <Direct> <Film/Snowden>} 구조로 올리버 스톤 감독이라는 개체에 대한 존재하는 사실들을 RDF 트리플로 표현한다.

1 AI Lab., Saltlux, Seoul, 06147, Korea
2 Dept. of Computer Engineering, Seokyeong University, Seoul, 02713, Korea
* Corresponding author (syl@skuniv.ac.kr)
[Received 4 March 2018, Reviewed 12 March 2018, Accepted 24 May 2018]



(그림 1) RDF 트리플 구성 예
(Figure 1) Example of RDF Triples

RDF 트리플의 각 요소는 URI로 식별되며 HTTP로 접근된다. 단, 목적어의 경우 리터럴 형식도 가능하다. 개체는 클래스로 정의되며 주어와 목적어는 이러한 클래스의 개체로 배정된다. (그림 1) 예에서 <Personal/OliverStone> 개체는 Person을 <Film/Snowden> 개체는 Film을 클래스로 하는 경우를 상정하였다. 술어는 공개된 어휘집에 자신의 역할과 정의역, 치역이 정의되어있다. 예를 들어 술어 <mov:Direct>에는 Person과 Film이 각각 정의역과 치역으로 정의되었다. 이러한 내용들은 LOD의 지식명세에 해당하며 이를 온톨로지라고 한다[5]. 온톨로지는 LOD와 함께 공개되어 제3자로 하여금 LOD 내용의 의미적 구조를 파악할 수 있도록 한다. LOD와 온톨로지는 W3C에서 공표한 RDF, RDFS, OWL, OWL2 언어로 기술된다[6]. LOD에 대한 질의는 W3C의 SPARQL 형식으로 구성되며 LOD는 질의를 접수하고 결과를 반환하는 SPARQL Endpoint 프로세스를 운영한다.

LOD 클라우드에서의 지식확장은 특정 개체에 대하여 여러 LOD들이 각자의 관점에서 기술한 내용을 획득함으로써 이루어진다. (그림 1)은 <owl:sameAs>를 이용하여 LODA의 <Personal/OliverStone>에 대한 지식확장을 이루는 예를 보이고 있다. 즉, LODA의 <Personal/OliverStone> 개체와 LODB의 <Movie/OliverStone> 개체가 동일하다는 사실을 {<Personal/OliverStone> <owl:sameAs> <Movie/OliverStone>}로 LODA에 기술함으로써 LODA의 <Personal/OliverStone>를 접근하여 얻은 {<OliverStone> <Homepage> <http://www.oliverstone.com>}, {<OliverStone> <Gender> <'Male'>}에 더하여 동일연결을 통하여 LODB로부터 획득한 내용을 {<OliverStone> <Direct> <Snowden>}, {<OliverStone> <Genre> <'Documentary'>}로 재구성하여 제공하는 것이 가능하여진다. 개체간 동일연결은 이렇듯 사용자로 하여금 예상하지 못하였던 주제의 내용으로 확장된 지식을 획득할 수 있도록 한다. 따라서, LOD 클라우드의 발전을 위해서는 동일연결을 확충하는 것이 필수적이다[7].

하지만, 오늘날의 LOD 클라우드는 지속적인 양적 성장[3]에도 불구하고 LOD간 연결이 매우 부족한 상태이다. LOD들의 44%는 데이터 사일로 수준으로 운영되고 있으며 2개 이하 LOD로만 연결된 경우도 71%에 이르고 있다[8]. 이렇듯 동일연결 확충이 지체되는 주요 원인은 동일연결 생성 자체가 어렵다는데 있다. LOD들은 평균 18,000여 개체와 67,000여 트리플들로 구성되어 있다[3]. 이러한 방대한 양의 LOD 내용을 파악하고 이들로부터 동일연결 시킬 개체쌍을 선별하는 것은 현실적으로 무척 어려운 작업이다. 이를 해결하기 위하여 선행연구들이 <owl:sameAs>로 기술된 동일연결 트리플들을 자동생성하고 이를 LOD에 첨부하여 공개하는 방안을 제시하였다. 하지만, 이 방안은 첨부된 동일연결들을 유지, 관리하는데 어려움을 발생시켰고 LOD의 갱신내용을 검색결과에 적시에 반영하지 못하는 문제점을 가진다[9].

이에 본 연구에서는 LOD 클라우드에서 수월하게 동일연결을 확충할 수 있는 방안으로 LOD마다 연결정책을 수립, 공개할 것을 제안하였다. 더불어, 이러한 연결정책들을 기반으로 여러 LOD들에 걸친 심층검색을 수행하고 그 결과를 일반 LOD와 동일한 형태로 재구성하여 제공하는 시스템(Policy based In-depth Searching and Cleansing: PISC)을 구현하였고 이를 Github를 통하여 공개하였다*. PISC는 검색한 개체들의 동일수준을 평가하고 이들을 사

* https://github.com/foxcats/LPS, 2018

용자가 요구한 동일기준 이상의 개체들만으로 정제하여 제공하는 기능도 함께 제공한다.

본 논문의 2장에서는 동일연결 생성 및 웹에 존재하는 개체들에 대한 신뢰수준 평가와 관련된 선행연구들을 고찰하였다. 3장에서는 PISC의 시스템 구성을 제시하고 각 요소의 운영과정을 설명하였으며 연결정책 수립을 위한 어휘집합도 함께 제시하였다. 4장에서는 PISC에 대한 실험을 수행하여 검색결과의 확장률과 포함률, 검색된 개체의 동일수준을 분석하였다. 5장에서는 향후 연구 방향을 제시하였다.

2. 관련연구

LOD 클라우드에서의 동일연결 생성방식은 크게 표준식별자 방식과 유사목적어 방식으로 나누어진다. 표준식별자 방식은 역함수 기능을 가지는 술어가 표준식별자를 목적어로 하는 환경에서 적용된다. 대표적인 표준식별자로는 ISBN(International Standard Book Number), GTIN(Global Trade Item Numbers), ISIN(International Securities Identification Numbering) 등이 있다. 주어 개체들이 동일한 표준식별자 값을 목적어로 하고 이에 역함수 술어가 적용될 때 해당 주어개체들을 동일연결된 것들로 판단한다. 이 방식은 동일연결의 높은 정확성을 제공하지만 그 적용이 표준식별자가 사용되는 경우로만 국한되는 한계를 가진다[4].

유사목적어 방식은 연결대상 LOD의 온톨로지를 분석하여 동일한 역할을 하는 술어쌍을 선정한 후 이러한 술어들로 연결된 목적어들이 유사한 경우 해당 주어 개체들을 동일연결 할 수 있는 것으로 판단한다. 필요할 경우 선정된 술어별로 평가 가중치를 부여한다. 이러한 방식에는 SILK[10], LIMES[11], SALE[12], TILE[13] 등 다수의 연구들이 진행되었다. SILK는 1차적으로 동일연결 개체 후보군을 구축한 후 이를 대상으로 미리 선정한 술어쌍들을 이용하여 유사성을 평가하는 방식으로 운영된다. LIMES는 SILK에 대한 성능향상에 중점을 두었다. SILK와 유사한 방식으로 동일연결 개체 후보군을 구축한 후 이들로부터 표본집합을 구성하고 각 표본과 가장 인접한 개체들을 공간위치를 이용한 삼각부등식 방식으로 선별한다. 이러한 과정에서 결코 유사할 수 없는 개체들을 사전에 제외시킴으로써 동일연결 개체 선별의 성능을 향상시킬 수 있었다.

SALE과 TILE은 SILK와 LIMES가 미리 선정한 술어쌍의 목적어 값에 국한하여 유사성을 평가하는데 더하여 술어가 가지는 문법적 특성을 추가적으로 반영하는 방식을 제시하였다. SALE은 RDFS, OWL을 문법점검 대상으로 하였고 TILE은 이에 OWL2를 확장하였다. OWL2가 포함됨에 따라 추론특성도 유사성 평가에 반영되었다. SILK, LIMES, SALE, TILE은 모두 소스와 타겟 개체쌍에 <owl:sameAs>를 적용하여 연결트리플을 생성하고 이들을 LOD와 함께 공개하는 방식을 채택하였다. 이러한 방식은 새로이 추가된 개체들이 검색결과에 반영되는 것은 이들에 대한 동일연결 트리플들이 LOD에 추가되기 전까지는 불가능하다는 한계를 가진다.

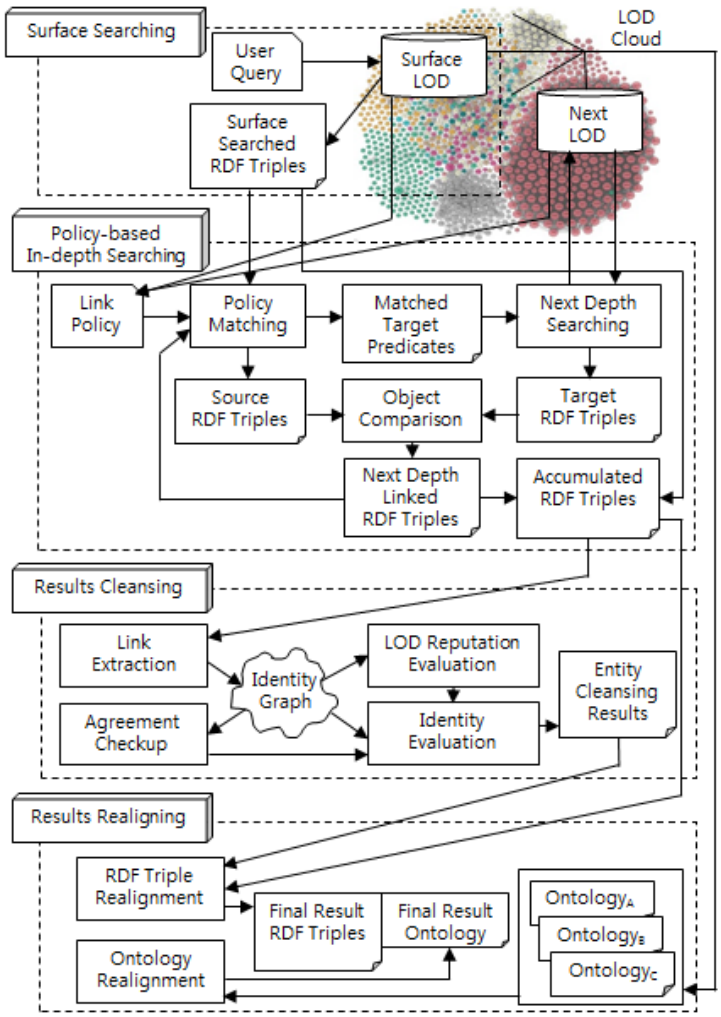
검색된 웹 페이지에 대한 신뢰평가와 관련된 대표적인 연구로는 구글의 페이지랭킹 알고리즘이 있다[14]. 이는 검색된 페이지를 참조하는 페이지들이 많을수록, 그리고 참조하는 페이지들의 신뢰수준이 높을수록 검색된 페이지의 신뢰수준을 높게 평가하는 방식이다. 본 연구에서도 심층 검색된 개체가 최초 검색된 개체와 동일하다는 것이 어느 정도의 신뢰를 가지는 것인가에 대한 평가에 페이지랭킹 방식과 유사한 정책을 적용하였다.

3. 연결정책기반 심층검색 및 정제시스템

사용자는 LOD 클라우드의 특정 LOD를 선정하여 질의를 요청한다. 이러한 질의에 대하여 PISC(Policy-based In-depth Searching and Cleansing)는 해당 LOD로부터 검색한 개체들에 대한 내용과 더불어 사용자가 인식하지 못하였던 LOD들로부터도 기 검색된 개체와 동일한 개체들을 검색하고 그 결과를 정제하여 사용자에게 반환하는 기능을 제공한다. (그림 2)는 PISC의 시스템 구성도를 도시한다. PISC는 표층검색 (Surface Searching), 연결정책기반 심층검색 (Policy-based In-depth Searching), 검색결과 정제(Results Cleansing), 검색결과 재구성(Result Realigning) 단계로 진행한다.

3.1 표층개체 검색

LOD 클라우드에 대한 사용자 접근은 특정 LOD를 표층LOD(Surface LOD)로 선정하는 것으로 시작한다. 우선 사용자는 PISC에 {SPARQL, SPARQL Endpoint, Depth Level, Similarity Level, Identity Level}을 제시한다. SPARQL은 RDF 데이터집합에 대한 W3C 표준 질의문법으로 작성된 질의문이다[15]. SPARQL Endpoint는 SPARQL을 받아들이고 검색결과를 제공하는 프로세스의



(그림 2) PISC 시스템 구성도
(Figure 2) System Architecture of PISC

URI로 표층LOD에서 제공하는 기능이다. Depth Level은 0 이상 정수이며 심층검색을 진행하는 깊이제한을 지정한다. 심층검색은 Depth Level이 1 이상인 경우에 진행한다. Similarity Level은 0.0 ~ 1.0 값을 가지며 소스와 타겟 RDF 트리플들의 술어쌍이 소스LOD의 연결정책에 명세된 경우 이들의 목적어들에 대한 유사 정도를 지정한다. 1.0인 경우 목적어들이 완전히 동일하다는 조건을 만족하여야만 소스와 타겟 주어개체들을 동일후보개체집합에 포함시킨다. Identity Level은 0.0 ~ 1.0 값을 가지는데 이는 여러 LOD들에 걸쳐 검색한 동일후보개체들 중 검색결과 최종 대상으로 선별할 때 적용하고자 하는 동일수준을 지정한다. 즉, 동일후보개체가 표층LOD에서 검색한 개체와 어느 수준 이상 동일하여야 하는가에 대한 기준점 역할을 한다. 이렇게 구성된 사용자 질의요구에 대하여 표층LOD로부터 검색된 결과는 표층검색트리플(Surface Searched RDF Triples)에 저장된다. 표층LOD는 사용자가 그 존재를 인식하여 질의를 요청한 대상이므로 충분히 신뢰적인 것으로 전제한다. 따라서, 표층검색결과는 모두 최종결과(Final Result RDF Triples, Final Result Ontology)에 포함된다.

3.2 연결정책 명세어휘

PISC의 심층검색은 연결정책을 기반으로 한다. 이에, 본 연구에서는 LOD마다 연결정책(Link Policy)을 수립하고 이를 LOD와 함께 공개함으로써 심층검색 과정에서 참조될 수 있는 환경을 마련할 것을 제안한다. 동일연결 트리플들을 미리 생성하여 LOD에 첨부하는 기존 방식은 첨부시점과 사용자 질의처리 시점간의 차이로 인하여 대상 LOD에서 추가된 개체의 내용이 검색결과에 포함되지 못하는 상황이 발생한다. 이러한 문제는 연결정책을 검색과정에서 참조하여 동일연결 개체 검색에 필요한 SPARQL을 구성함으로써 해결할 수 있다. 본 연구에서는 연결정책을 명세하는데 필요한 어휘들을 표 1과 같이 제안한다.

LOD(SourceLOD)는 lp:linkpolicy를 사용하여 다수의 상세 연결정책들을 명세한다. lp:linkpolicy는 목적어로 공백노드 (BlankNode_A)를 가지며 이 공백노드는 SourceLOD 연결정책 명세의 출발점이 된다. BlankNode_A의 술어 lp:targetLOD는 SourceLOD와 연결시킬 타겟LOD의 SPARQL EndPoint 주소를 등록한다. 한정주제 등록은 lp:regRestrictTopic으로 명세되며 목적어로 공백노드 (BlankNode_B)를 가진다. BlankNode_B는 lp:sourceRestric Predicate와 lp:sourceRestricTopic을 이용하여 소스LOD에서 한정주제에 사용될 술어를 명세한다. 타겟LOD에 대한 한정주제 등록은 lp:regTargetRestrict로 명세하며 목적어로 공백노드(BlankNode_C)를 가지며 여기에 lp:target RestricPredicate와 lp:targetRestricTopic을 사용하여 타겟 LOD에 적용할 한정주제 술어와 주제를 명세한다.

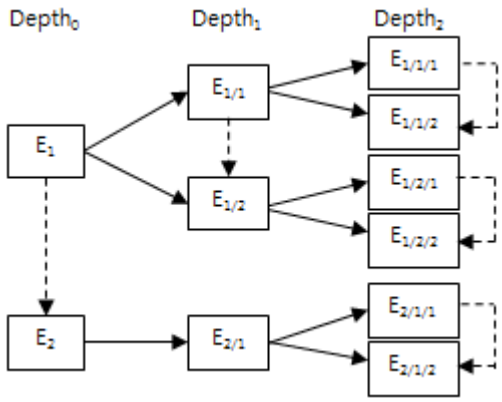
술어매칭은 lp:predicateMatching으로 명세한다. 술어매칭의 경우 한정주제에 귀속되어야 하므로 주어는 lp:reg RestrictTopic의 목적어인 BlankNode_B가 된다. lp:predicate Matching은 목적어로 공백노드(BlankNode_D)를 가지며 이를 공동주어로 하는 lp:sourcePredicate과 lp:target Predicate를 이용하여 소스LOD와 타겟LOD에서 동일한 의미를 가지는 매칭술어 쌍을 명세한다.

3.3 연결정책기반 심층검색

사용자 질의요구에서 Depth Level이 1 이상인 경우 연결정책기반 심층검색(Policy-based In-depth Searching) 단계를 진행한다. 연결정책매칭(Policy Matching)은 최초에는 표층검색트리플로부터 표층LOD의 연결정책에 명세된 소스술어가 포함된 RDF 트리플들을 발췌하여 소스트리플(Source RDF Triples)을 구성한다. 또한 연결정책매칭은 표층LOD의 연결정책을 참조하여 소스트리플에 포함된 술어와 매칭되는 타겟술어들을 선별하고 이를 해당 타겟LOD와 {한정주제, 타겟술어, 타겟LOD_URI}로 짝을 이루어 매칭타겟술어(Matched Target Predicates)에 등록한다. 한정주제는 타겟술어를 적용하여야 하는 주어개체의 주제를 한정함으로써 타겟LOD에서의 검색대상을 축소

(표 1) 연결정책 명세 어휘
(Table 1) Vocabularies of Link Policy Specification

Vocabulary	Role	Subject	Object
lp:linkpolicy	Link policy registration	SourceLOD	BlankNode_A
lp:targetLOD	Target LOD registration	BlankNode_A	Target LOD
lp:regRestrictTopic	Topic restriction registration	BlankNode_A	BlankNode_B
lp:sourceRestricPredicate	Predicate for source LOD's topic restriction registration	BlankNode_B	Predicate for source LOD's topic restriction
lp:sourceRestricTopic	Type for source LOD's topic restriction registration	BlankNode_B	Type for source LOD's topic restriction
lp:regTargetRestrict	Target LOD's topic restriction registration	BlankNode_B	BlankNode_C
lp:targetRestricPredicate	Predicate for target LOD's topic restriction registration	BlankNode_C	Predicate for target LOD's topic restriction
lp:targetRestricTopic	Type for target LOD's topic restriction registration	BlankNode_C	Type for target LOD's topic restriction
lp:predicateMatching	Predicate matching registration	BlankNode_B	BlankNode_D
lp:sourcePredicate	Source LOD's predicate registration	BlankNode_D	Source LOD's predicate
lp:targetPredicate	Target LOD's predicate registration	BlankNode_D	Target LOD's predicate



(그림 3) 연결정책기반 심층검색 진행 예
(Figure 3) Example of Policy-based In-depth Searching

시키는 역할을 한다. 차층검색(Next Depth Searching)은 매칭타겟술어가 제공하는 한정주제와 타겟술어를 조건으로 하는 SPARQL을 구성하고 이를 타겟LOD_URI에 근거하여 구성한 SPARQL Endpoint로 요청한다.

타겟LOD(Next LOD)로부터 검색한 결과는 타겟트리플(Target RDF Triples)에 저장한다. 목적어비교(Object Comparison)은 소스트리플과 타겟트리플을 대상으로 연결정책에 명세된 술어쌍에 부합하는 트리플쌍을 선별하고 이들의 목적어들이 사용자가 제시한 유사수준(Similarity Level) 이상인가를 확인한다. 목적어간 유사수준에 대한 평가에는 내부 문자열의 순서가 달라도 적정한 보정을 통하여 보다 정밀한 평가를 할 수 있는 N-gram 거리측정 방식을 적용하였다.

목적어가 충분히 유사한 것으로 확인된 타겟트리플은 차층연결트리플(Next Depth Linked RDF Triples)에 저장된다. 차층연결트리플은 해당 깊이에서 검색된 트리플들 가운데 표층검색 개체와 동일하다고 평가된 후보들이므로 이들을 누적트리플(Accumulated RDF Triples)에 누적시켜 다음 단계인 검색결과 정제를 위한 입력으로 준비한다. 차층연결트리플은 전술한 표층검색트리플과 마찬가지로 연결정책매칭의 입력으로 제공된다. 차층연결트리플을 제공하였던 LOD(Next LOD)에 명세되어 있던 연결정책은 이번 깊이에서의 연결정책(Link Policy)이 되고 연결정책매칭은 이를 참조하여 차층연결트리플로부터 소스트리플과 매칭타겟술어를 생성하여 전술한 방식의 이후 과정들을 진행한다. 이러한 진행은 사용자가 제시한 Depth Level에 도달할 때까지 이루어진다.

심층검색 과정에서의 개체들에 대한 접근순서는 깊이우선탐색으로 진행된다. 즉, (그림 3) 예에서 개체들에 대한 접근은 $E_1 \rightarrow E_{1/1} \rightarrow E_{1/1/1} \rightarrow E_{1/1/2} \rightarrow E_{1/2} \rightarrow E_{1/2/1} \rightarrow E_{1/2/2}$으로 진행된 후 $E_2 \rightarrow E_{2/1} \rightarrow E_{2/1/1} \rightarrow E_{2/1/2}$로 진행된다. $E_{1/1}$, $E_{1/1/1}$, $E_{1/1/2}$, $E_{1/2}$, $E_{1/2/1}$, $E_{1/2/2}$은 모두 개체 E_1에 대한, $E_{2/1}$, $E_{2/1/1}$, $E_{2/1/2}$는 개체 E_2에 대한 동일후보개체들이다.

3.4 검색결과 정제

심층검색 결과인 누적트리플에는 LOD간 연결정책에 의거하여 유사수준(Similarity Level) 이상으로 동일한 것으로 평가된 후보개체들에 대한 RDF 트리플들이 기록되어 있다. 이들 개체들 가운데 사용자가 제시한 동일수준(Identity Level) 이상으로 표층검색 개체들과 동일하다고 평가되는 것들을 정제해내는 과정을 검색결과정제(Results Cleansing) 단계에서 수행한다. 동일연결추출(Link Extraction)은 누적트리플 집합으로부터 각 표층개체 단위로 개체간 동일연결을 파악하여 동일그래프(Identity Graph)를 구축한다. (그림 4)의 동일그래프에서 정점은 (LOD, 개체)가 되며 방향성 간선은 동일평가에 적용된 술어쌍(P_A/P_B)을 가진다. 즉, "LOD L_1의 개체 E_1은 L_1의 술어 P_1과 L_2의 술어 P_2을 술어쌍으로 하여 LOD L_2의 개체 E_2와 동일한 것으로 후보등록 되었다"라는 사실이 동일그래프의 한 부분으로 기술되었다. E_1, E_2, E_3, …, E_i는 E_0와 동일후보 개체들이지만 소스개체 E_1과 타겟개체 E_2간 동일연결은 소스LOD L_1의 연결정책을 심층검색에 적용한 결과이다. 즉, L_1의 연결정책은 L_2의 술어 P_2는 인식하고 있지만 L_3의 술어 P_3는 인식하지 않고 있다. 따라서, 동일그래프의 시작점인 표층개체로부터 깊이를 더하여 멀어질수록 표층개체와의 동일수준은 낮아질 것이다. 즉, E_0와 동일하다는 수준은 E_1보다 E_i가 더 낮은 것으로 평가하는 것이 타당할 것이다.

동일그래프를 (그림 4)에서는 단순한 예로 제시하였지만 실제로는 정점으로 다수의 진입간선들이 존재한다. 진입간선이 많다는 것은 다른 LOD들로부터 해당 LOD로의 동일연결 요청이 많다는 것이고 이는 해당 LOD에 대한 다른 LOD들로부터의 평판이 높다는 것으로 해석할 수 있다. 구글의 페이지랭킹 알고리즘은 페이지$_A$를 참조하는 페이지들이 많을수록, 즉 진입연결이 많을수록, 그리고 참조하는 페이지의 신뢰도가 높을수록 페이지$_A$의 신뢰도를 높게 평가하는 것을 기본 정책으로 하고 있다. PISC의 LOD평판평가(LOD Reputation

Evaluation)에서도 이와 유사한 방식으로 평가한다. 단, 페이지 별로 평판을 미리 평가한 결과를 사용하는 구글 페이지랭킹과는 달리 LOD평판 평가는 각 질의단위로 이루어진다. 사용자 질의를 처리하는 과정에 참여하였던 모든 LOD들과 개체들을 동일그래프가 포함하고 있으므로 질의단위로 LOD평판을 (식 1)로 평가하는 것이 가능하다.

(식 1) LOD평판 평가

$Rep(LOD_{cur}) = ((InC_{cur} - InC_{min})/(InC_{max} - InC_{min})) * 0.3 + 0.7$

- InC_{max}: 가장 많은 진입동일연결을 가지는 LOD의 진입동일연결 수
- InC_{min}: 가장 적은 진입동일연결을 가지는 LOD의 진입동일연결 수
- InC_{cur}: 현재 평가대상인 LOD의 진입동일연결 수 ■

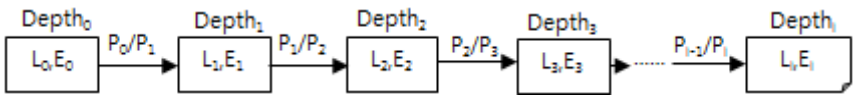
(식 1)을 운용함에 따라 LOD평판은 0.7 ~ 1.0 값을 가진다. 이는 LOD를 구축하고 이를 공개하는 기관들이 일정수준 이상의 공신력을 가지는 것으로 전제하는 것이 현실적이라는 판단에 근거한다. 하지만, 표층LOD인 LOD_0는 평판을 1.0으로 간주하였다. 사용자가 제시한 동일수준(Identity Level) 이상의 개체들을 동일개체 후보들로부터 선별하기 위하여 동일수준평가(Identity Evaluation)는 이들이 표층검색 개체 E_0와 동일한 정도를 0.0 ~ 1.0 값으로 평가한다. 동일그래프를 구성하는 $(LOD_i, E_i) \rightarrow (LOD_j, E_j)$에 대하여 개체 E_j의 동일수준 $Ident(E_j)$는 (식 2)로 평가한다.

(식 2) 동일수준 평가

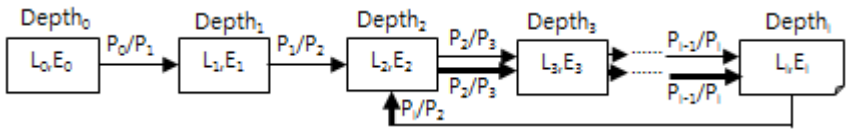
$Ident(E_j) = Rep(LOD_i) \times Ident(E_i)$ ■

$Ident(E_j)$는 개체 E_i가 표층검색 개체 E_0와 동일한 정도에 연결정책을 통하여 E_j를 향한 동일연결을 생성한 LOD_i의 평판을 반영한 결과가 된다. 따라서 평판이 높은 LOD로부터 동일연결된 개체의 동일수준은 높게 평가된다. LOD_0의 평판을 1.0으로 전제한 것과 마찬가지로 표층개체 E_0의 동일수준도 1.0으로 전제하였다. 이는 동일수준 기준점이 E_0이므로 타당한 전제로 판단된다. (식 2)를 운영함에 따라 동일수준은 동일그래프의 간선을 따라 전파된다. 이에 따라 여러 LOD들을 거쳐 심층검색을 진행할수록 동일개체 후보들의 동일수준은 감소하는 추이를 보이게 된다. 더불어, $(LOD_{i1}, E_{i1}) \rightarrow (LOD_j, E_j)$, $(LOD_{i2}, E_{i2}) \rightarrow (LOD_j, E_j)$와 같이 개체 E_j에 2개 이상의 동일연결들이 진입할 경우 $Ident(E_j)$는 이들 가운데 가장 큰 값을 자신의 동일수준으로 채택한다.

동일동조점검(Agreement Checkup)에서는 동일그래프에서 사이클이 발생하였는가를 점검하고 동일수준평가로 하여금 이를 반영하도록 하는 역할을 한다. (그림 5)는 개체 E_i로부터 E_2가 궁극적으로 E_0에 대한 동일후보라는 사실에 대한 동조가 발생하는 예를 도시하고 있다. 즉, $E_0 \rightarrow E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow \ldots \rightarrow E_i$로부터 E_i가 E_0와 동일후보라는 사실이 추론되고 그러한 E_i로부터 E_2를 향한 동일연결이 이루어졌다. 동일동조 발생은 동일수준평가에 전달되어 E_2의 동일수준은 기존의 동일수준 결과값에 $Rep(LOD_i) \times Iden(E_i)$가 더하여진 결과로 수정된다. 동일동조 사이클에 포함된 나머지 개체 E_3, …, E_i의 동일수준도 동일한 방식으로 동일수준평가를



(그림 4) 동일그래프 예
(Figure 4) Example of Identity Graph



(그림 5) 동일동조 발생 예
(Figure 5) Example of Identity Agreement

통하여 보강된다. 단, 보강된 동일수준이 1.0을 초과할 경우 1.0으로 한다. 동일그래프에 등장하는 모든 동일 후보 개체들에 대한 동일수준 평가가 완료하면 이들 가운데 사용자가 제시한 동일수준(Identity Level)을 상회하는 개체들만을 개체정제결과(Entity Cleansing Results)에 수록하고 이를 검색결과재구성 단계에 전달한다.

3.5 검색결과 재구성

검색결과 재구성(Results Realigning) 단계에서는 최종 검색결과를 사용자로 하여금 RDF 모델에 입각하여 활용할 수 있도록 재구성한다. 이를 위하여 트리플재구성(RDF Triple Realignment), 온톨로지재구성(Ontology Realignment)를 진행하고 최종결과로 검색결과(Final Searching Results)와 재구성된 온톨로지(Ontology Realigned)를 쌍으로 하여 제공한다. 트리플재구성은 개체정제결과를 참조하여 E0와 동일수준 이상으로 동일한 것으로 선별된 개체들에 대한 상세 내용들을 누적트리플에서 가져와 E0를 주어로 하는 {E0, 술어, 목적어} 형식의 RDF트리플 집합으로 재구성하여 검색결과에 추가한다. 사용자는 질의 시 표층LOD만을 인식하였으므로 이로부터 검색된 표층개체를 주어로 하여 여러 LOD들로부터 획득한 다른 관점에서 기술된 술어, 목적어 쌍들로 RDF 트리플을 구성하였다.

온톨로지재구성(Ontology Realignment)는 검색결과에 대한 의미구조를 제공함으로써 사용자로 하여금 검색결과 내용의 의미를 파악하는 것을 가능하게 한다. 이를 위하여 검색결과에 포함된 개체가 있었던 LOD의 온톨로지를 접근하여 개체 기술에 사용된 술어의 문법적 특성과 용도, 관련 클래스 정의를 수집하고 이를 검색결과에 대한 온톨로지(Ontology Realigned)로 재구성하는 작업을 진행한다. 트리플재구성과 온톨로지재구성 결과인 검색결과와 재구성된 온톨로지를 PISC는 N-Triple 방식으로 직렬화하였다.

4. 실험 및 분석

PISC에 대한 실험은 DBpedia에 참여하고 있는 한국, 이탈리아, 프랑스, 포르투갈, 스페인의 LOD를 대상으로 진행하였다. 각 LOD의 규모는 표 2와 같다. SPARQL 처리에는 Apache Jena 3.1.0 API가 사용되었다. 연결정책에 명세된 술어에 연결된 목적어들간의 유사도는 1.0, 0.9, 0.8, 0.75를 적용하여 진행하였다. 한국LOD를 표층LOD로 하여 '아이언맨, '분노의질주', '스파이더맨', '배트맨'을 검색조건으로 SPARQL을 요청하였다. 각 LOD별로 2, 3, 4개 LOD를 타겟LOD로 정책에 명세하는 경우들을 상정하였으며 심층검색은 최대 깊이4까지 진행하였다. 분석은 검색결과의 확장률, 포함률, 동일수준을 대상으로 하였다.

(표 2) 실험 LOD 규모
(Table 2) Scale of LODs

LOD	개체 수
http://ko.dbpedia.org	310,811
http://fr.dbpedia.org	1,591,318
http://it.dbpedia.org	968,794
http://es.dbpedia.org	1,120,144
http://pt.dbpedia.org	865,889

4.1 확장률 분석

확장률은 PISC가 연결정책을 이용하여 심층검색을 진행할 경우 <owl:sameAs>만으로 검색되는 개체수와 비교하여 어느 정도 비율로 검색하는가를 나타낸다. 확장률 계산에는 (식 3)이 적용된다.

(식 3) 깊이$_i$에서의 확장률

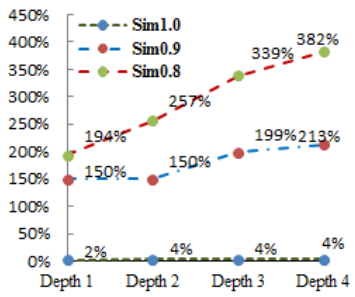
$Exp(i) = (Policy(i) - SameAs(i))/SameAs(i)$

- Policy(i): PISC가 깊이$_i$까지 검색한 개체수
- SameAs(i): <owl:sameAs>만으로 깊이$_i$까지 검색한 개체수

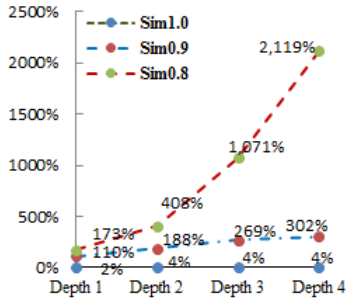
단, Policy(i)와 SameAs(i)에 참여한 LOD들은 동일함 ■

(그림 6), (그림 7), (그림 8)은 타겟LOD가 2개, 3개, 4개로 명세된 연결정책을 이용하여 심층검색한 결과 개체들에 대한 평균 확장률을 나타낸다. Sim1.0, 0.9, 0,8은 매칭술어에 연결된 목적어들의 유사도 1.0, 0.9, 0.8에 해당한다. Sim0.75의 경우 표층LOD에서 검색된 개체와 과도하게 상이한 개체들이 대량으로 검색됨에 따라 그래프 표기에서 제외하였다.

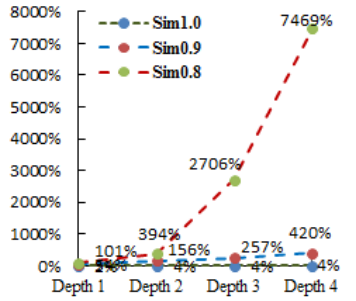
Sim1.0인 경우 타겟LOD수와 무관하게 <owl:sameAs>로만 검색한 결과와 비교하여 확장이 거의 이루어지지 않았다. Sim0.9, Sim0.8 경우 깊이를 더할수록 확장률이 증가하였다. 하지만 타겟LOD가 3개와 4개인 경우 깊이 3, 4에서 Sim0.8은 Sim0.9와 비교하여 과도한 확장을 보이고 있다. 타겟LOD가 2개인 경우에서는 Sim0.8, Sim0.9에서 적절한 확장이 이루어졌다. 하지만 연결정



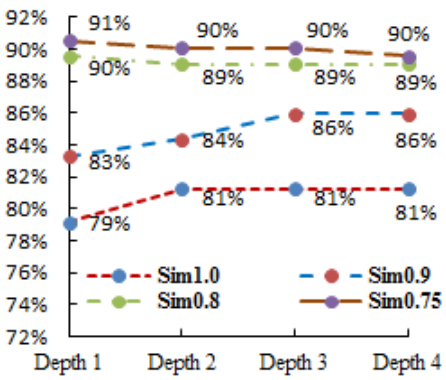
(그림 6) 타겟LOD 2개에서의 확장률
(Figure 6) Expansion ratio of 2 Target LODs



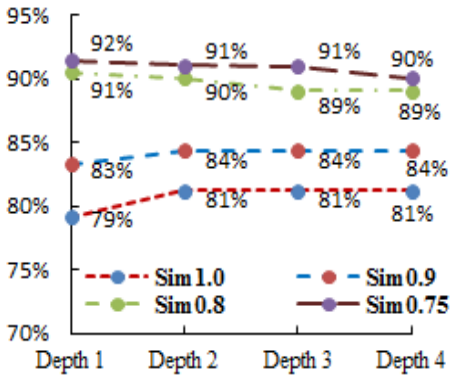
(그림 7) 타겟LOD 3개에서의 확장률
(Figure 7) Expansion ratio of 3 Target LODs



(그림 8) 타겟LOD 4개에서의 확장률
(Figure 8) Expansion ratio of 4 Target LODs



(그림 9) 타겟LOD 2개에서의 포함률
(Figure 9) Inclusion ratio of 2 Target LODs



(그림 10) 타겟LOD 3개에서의 포함률
(Figure 10) Inclusion ratio of 3 Target LODs

책에 등장하는 타겟LOD수가 2개 이상인 경우가 일반적이므로 확장률 측면에서는 목적어의 유사도를 0.9 정도로 적용하는 것이 적절할 것으로 판단된다.

4.2 포함률 분석

포함률은 심층검색을 진행하였던 LOD들로부터 검색한 개체들에 동일한 LOD로부터 <owl:sameAs>만으로 검색한 개체들이 얼마나 포함되어 있는가를 나타낸다. 이 실험에서는 DBpedia와 같이 연결이 매우 신뢰적으로 이루어진 LOD들을 대상으로 진행한 PISC의 검색결과가 이러한 신뢰적인 결과를 얼마나 포함할 수 있는가를 점검한다. 포함률 계산에는 (식 4)를 적용한다.

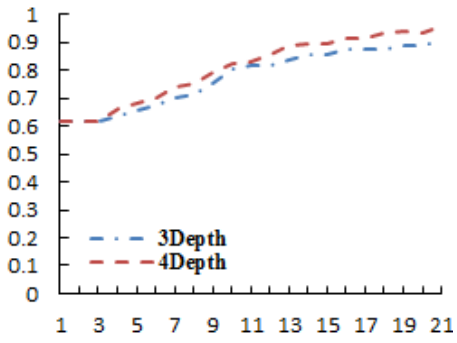
(식 4) 깊이$_i$에서의 포함률

$$Inc(i) = (Policy(i) \cap SameAs(i))/SameAs(i) \quad \blacksquare$$

(그림 9), (그림 10)는 타겟LOD가 2개, 3개로 명세된 연결정책으로 심층검색한 개체들의 평균 포함률을 나타낸다. 유사도는 Sim1.0, 0.9, 0.8, 0.75를 적용하였다. Sim0.75와 Sim0.8의 경우 높은 포함율을 보이고 있지만 전술한 바와 같이 과도한 확장률로 실제 적용하기에 어려움이 있다. Sim0.9, 1.0의 경우 Sim0.8 보다는 조금 낮지만 적절한 포함률을 제시하고 있다. 따라서, 포함률과 확장률을 함께 고려한다면 유사도는 0.9 정도를 적용하는 것이 적절할 것으로 판단된다.

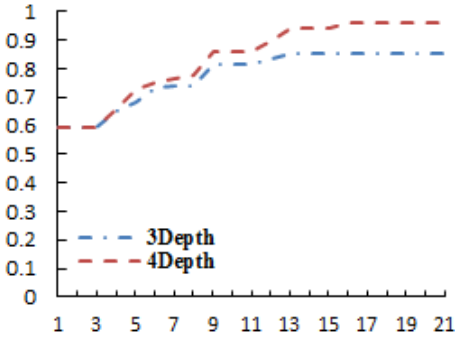
4.3 동일수준 분석

PISC 수행에 따른 동일수준의 변화에 대한 실험결과는 (그림 11), (그림 12), (그림 13)에서 제시하였다. 그림에서 세로축은 동일수준, 가로축은 동일동조 횟수를 나타낸다. 동일동조 횟수란 (그림 5)의 예와 같이 어떤 개체에



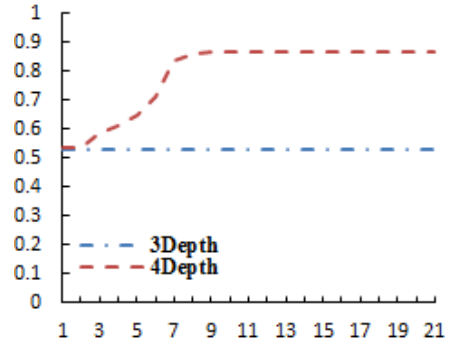
(그림 11) 타겟LOD 2개에서의 동일수준

(Figure 11) Identity Level of 2 Target LODs



(그림 12) 타겟LOD 3개에서의 동일수준

(Figure 12) Identity Level of 3 Target LODs



(그림 13) 타겟LOD 4개에서의 동일수준

(Figure 13) Identity Level of 4 Target LODs

대한 동일그래프에서 발생한 동일동조들을 누산한 값이다. 실험은 타겟LOD가 2개, 3개, 4개인 경우와 함께 깊이를 3과 4로 진행한 경우의 조합으로 진행하였다. 따라서 실험결과 값은 해당 깊이와 해당 동일동조 횟수에서의 동일수준값을 나타낸다. 그림에서 제시된 동일수준값들은 검색된 개체들의 동일수준값들에 대한 동일동조횟수별 평균값이다. 평균을 적용함에 따라 실험결과에 동일수준 1에 도달하는 경우가 등장하지 않지만 실제로는 동일수준 1에 도달하는 경우들이 존재한다. 보다 적은 동일동조횟수에서 보다 높은 동일수준값을 나타낼수록 사용자는 충분히 신뢰적인 것으로 평가된 개체를 보다 이른 시점에서 획득할 수 있게 된다.

동일수준에 대한 실험결과 연결정책에 타겟LOD를 3개, 4개를 명세하는 경우가 2개를 명세하는 경우보다 높은 동일수준을 얻을 수 있음을 확인할 수 있었다. 특히, 타겟LOD가 2개인 (그림 12)의 경우 깊이4에서도 동일수준이 1에 도달하는 경우가 매우 적었다. 깊이3인 경우에는 모든 개체들이 많은 동일동조에도 불구하고 동일수준 1에 도달하지 못하는 결과를 보였다. 따라서 사용자가 동일수준을 0.9 이상으로 요구할 경우 타겟LOD가 2개만 연결정책에 명세되어 있었다면 심층검색된 개체들의 대부분은 검색결과정제 단계를 통과하지 못하게 된다. 따라서 연결정책에는 타겟LOD를 3개 이상 명세하는 것이 동일성에 대한 충분한 검증이 이루어진 개체들을 다수 확보할 수 있는 방안이 될 것으로 판단된다.

5. 결론 및 향후 연구

본 연구에서는 LOD 클라우드가 지향하고 있는 지식의 확장을 위하여 LOD별로 연결정책을 수립, 공개할 것을 제안하였다. <owl:sameAs>를 이용한 기존의 개체간 동일연결 제공 방식은 구축과 운영상에 어려움이 야기한다. LOD마다 연결정책을 운영할 경우 동일연결성을 효과적으로 확충할 수가 있고 추가되었던 개체들을 검색시점에서 결과에 반영할 수 있다. 본 연구에서는 연결정책 지원 환경에서 심층검색을 수행하고 그 결과를 사용자가 제시한 동일수준 이상으로 정제하여 제공하는 PISC 시스템을 구현하였다. 실험은 DBpedia LOD를 대상으로 진행하였으며 연결정책에 명세할 목적어 유사도는 0.9 정도가 적절하고 연결정책에는 3개 이상의 타겟LOD들을 명세하는 것이 바람직 한 것으로 파악되었다.

향후 연구에서는 연결정책에 명세되는 술어들의 문법적 특징을 개체동일성 평가에 반영하는 방안을 마련하고자 한다. 특히 OWL2의 문법적 특징을 반영할 경우 RDF 모델이 가지고 있는 추론특성을 적극적으로 활용할 수 있을 것으로 예상된다. 또한, 연결정책 명세를 위한 어휘를 보강하여 보다 구체적인 연결정책 수립이 가능하도록 하는 방안을 마련하고자 한다.

참고문헌(References)

[1] Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, pp. 56-71, 2011

[2] T. B. Lee, “Semantic Web Road Map”, https://www.w3.org/DesignIssues/Semantic.html, 1998

[3] A. Abele and J. McCrae, “The Linked Open Data cloud diagram, 2017”. http://lod-cloud.net/, 2018

[4] Harth, A., et al., Linked Data Management, 1st Ed., 20-25. CRC Press, pp. 31-68, 2014

[5] Heath, T. and Bizer, C. Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, pp. 178-220, 2011

[6] A. Dean and H. James, Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, Elsevier, pp 132-143, 2011

[7] N. Konstantinou, N., Materializing the Web of Linked Data, 1st Ed., Springer, pp. 118-132., 2015

[8] C. Bizer, “Is the Semantic Web what we expected, 2017”. https://www.slideshare.net/bizer/is-the-semantic-web-what-we-expected-adoption-patterns-and-contentdriven-challenges-iswc-2016-keynote, 2016

[9] W3C, “What is Linked Data, 2017”, https://www.w3.org/standards/semanticweb/data, 2017

[10] J. Volz J., et al., “Silk – A Link Discovery Framework for the Web of Data”, Proc. of the 2nd Workshop on Linked Data on the Web 2009, pp. 238-247, 2009. http://www.researchgate.net/publication/228638267_Silk-A_Link_Discovery_Framework_for_the_Web_of_Data

[11] A. Ngonga and S. Auer, “LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data”, Proc. of the 22nd IJCAI, pp. 2312-2317, 2011. http://svn.aksw.org/papers/2011/WWW_LIMES/public.pdf

[12] J. Park and Y. Sohn., “A Syntax Added Link Evaluation Technique for Improving Trustworthiness of LOD’s Linkages”, Journal of KIISE: Databases, Vol. 41, No. 1), pp. 45-61, 2014. http://www.dbpia.co.kr/Journal/ArticleDetail/NODE02360287

[13] J. Park and Y. Sohn., “Trustworthiness Improving Link Evaluation Technique for LOD Linkages giving Considerations to the Syntactic Properties of RDFS, OWL, and OWL2”, Journal of KIISE: Databases, Vol. 41, No. 4, pp. 226-241, 2014. http://www.dbpia.co.kr/Journal/ArticleDetail/NODE02457716

[14] S. Brin, et al., “The PageRank Citation Ranking-Bringing Order to the Web”, http:// ilpubs.stanford.edu/422/1/1999-66.pdf. 1998

◐ 저 자 소 개 ◑

김 광 민(Kwangmin Kim)
2016년 서경대학교 컴퓨터공학과(공학사)
2018년 서경대학교 대학원 전자및컴퓨터공학과(공학석사)
2018년~현재 ㈜솔트룩스 인공지능연구센터
관심분야 : Linked Open Data, 온톨로지, 머신러닝, 빅데이터
E-mail: ip9894@naver.com

손 용 락(Yonglak Sohn)
1986년 경북대학교 전자공학과 전산전공(공학사)
1988년 고려대학교 대학원 전자및전산학과(공학석사)
2000년 KAIST 대학원 정보및통신공학과(공학박사)
1995년~현재 서경대학교 컴퓨터공학과 교수
관심분야 : 트랜잭션관리, 데이터모델링, 온톨로지, Linked Open Data, 정보보안
E-mail: syl@skuniv..ac.kr