KIPS Tr. Software and Data Eng.
Vol.3, No.12 pp.523~530 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.12.523

Ontology Alignment by Using Discrete Cuckoo Search

Jun-Han† · Hyunjun Jung†† · Doo-Kwon Baik†††

ABSTRACT

Ontology alignment is the way to share and reuse of ontology knowledge. Because of the ambiguity of concept, most ontology alignment systems combine a set of various measures and complete enumeration to provide the satisfactory result. However, calculating process becomes more complex and required time increases exponentially since the number of concept increases, more errors can appear at the same time. Lately the focus is on meta-matching using the heuristic algorithm. Existing meta-matching system tune extra parameter and it causes complex calculating, as a consequence, the results in the various data of specific domain are not good performed. In this paper, we propose a high performance algorithm by using DCS that can solve ontology alignment through simple process. It provides an efficient search strategy according to distribution of Levy Flight. In order to evaluate the approach, benchmark data from the OAEI 2012 is employed. Through the comparison of the quality of the alignments which uses DCS with state of the art ontology matching systems.

Keywords : Ontology, Ontology Alignment, Cuckoo Search

이산 Cuckoo Search를 이용한 온톨로지 정렬

한 군† · 정 현 준†† · 백 두 권†††

요 약

온톨로지 정렬의 목적은 지식을 공유 및 재사용 하는 데 있다. 기존 온톨로지 정렬 시스템은 온톨로지 개념의 모호성 때문에 여러 가지 다양한 측정 기법을 사용하고 전수조사를 수행하여 사용자가 만족하는 결과를 얻는다. 온톨로지 개념이 점차 많아짐에 따라 계산이 복잡해지고 걸리는 시간이 기하급수적으로 증가하여 처리 과정에서 오류가 발생한다. 이를 해결하기 위하여 메타 휴리스틱 알고리즘을 사용하는 메타 매칭이 연구되고 있다. 기존 메타 매칭 시스템에서는 사용하는 파라미터가 많기 때문에 온톨로지 정렬 처리에 계산이 복잡하고 특정 도메인의 다양한 데이터에 따라 조율이 요구되어 온톨로지 정렬 탐색에 좋은 성능을 보여주지 못했다. 이 논문에서는 온톨로지 정렬을 쉽고 간단한 계산을 통해 높은 성능을 목표로 하여 DCS(Discrete Cuckoo Search) 를 사용한 온톨로지 정렬 알고리즘을 제안한다. 제안한 알고리즘은 Levy Flight 분포에 따른 탐색으로 효율적인 전략을 보여준다. 제안된 알고리즘은 OAEI 2012(Ontology Alignment Evaluation Initiative)에서 제공하는 벤치마크 데이터와 제안 알고리즘을 사용하여 성능을 평가한다.

키워드 : 온톨로지, 온톨로지 정렬, Cuckoo Search

1. 서 론

최근 W3C에서는 기존의 인간이 중심이었던 웹으로부터 컴퓨터가 중심인 시맨틱 웹을 차세대 웹으로 연구하고 있다. 시맨틱 웹은 기존 웹을 확장하여 웹 상의 각 정보에 의미를 부여하여 컴퓨터가 지능적으로 정보 자원을 읽고 해석하여 자동으로 사용자의 요구에 따라 처리한다. 시맨틱 웹은 온톨로지를 활용하여 서비스를 기술하고 공유된 개념 의미 표현을 이용하여 서비스 관한 처리를 자동화한다[1]. 온톨로지는 인공지능, 정보시스템, 데이터 통합, 정보 검색, 지식관리 등 분야에서 활발히 응용 및 개발되고 있다. 대부분의 온톨로지는 같은 도메인을 다루더라도 작성된 지역 또는 전문가의 주관성에 의해 표현한 개념 형식이 달라 이질성이 발생한다. 이는 특정 도메인의 온톨로지 지식 공유 및 재사용에 영향을 미치고 온톨로지 간의 상호운용성의 병목현상으로 간주된다. 분산된 이질적인 환경에서 온톨로지 서비스 공유를 실현하기 위하여 온톨로지들 간에 의미 기반 대응관계를

※ 이 논문은 2014년 정부(미래창조과학부)의 재원으로 한국연구재단 차세대 정보.컴퓨팅 기술개발사업의 지원을 받아 수행된 연구임 (NRF-2012M3C4A7033346).
† 준 회 원 : 고려대학교 컴퓨터 · 전파통신공학과 석사과정
†† 준 회 원 : 고려대학교 컴퓨터 · 전파통신공학과 박사과정
††† 종신회원 : 고려대학교 컴퓨터 · 전파통신공학과 교수
Manuscript Received : August 6, 2014
First Revision : September 23, 2014
Accepted : October 1, 2014
* Corresponding Author : Doo-Kwon Baik(baikdk@korea.ac.kr)

이루는 연관된 개념들을 찾는 온톨로지 매칭 기법이 연구되고 있다. 이를 기반하여 대응관계들의 집합을 이루는 온톨로지 정렬 시스템이 다양하게 개발되고 있다.

온톨로지 매칭은 유사도 측정 기법을 사용하여 구한 유사도 값에 의해 결정된다. 단일 유사도 측정 기법 시스템에서는 온톨로지 개념의 이질성과 표현의 모호성으로 인해 사용자가 원하는 결과를 얻기 힘들다. 최근 온톨로지 정렬 시스템은 여러 개의 유사도 측정 기법을 결합하여 서로 다른 정보 타입 (label, text, description, structure, rules)을 각각 처리하고 각 항목의 유사도 값을 가중치 조합에 따라 계산하여 결과의 정확성을 높였다[2]. 그러나 기존 복합시스템에서는 온톨로지 개념의 양에 따라 매트릭스를 생성하고 전수조사를 진행하여 온톨로지 정렬을 수행한다. 이는 온톨로지 개념의 양의 규모가 커짐에 따라 계산이 복잡해지고 걸리는 시간이 기하급수적으로 증가되며 오류가 발생빈도가 증가한다. 이러한 문제를 해결하기 위하여 최근에는 메타 매칭 기법을 사용하는 연구가 대두되고 있다.

메타 매칭은 메타 휴리스틱 알고리즘을 사용하여 온톨로지 정렬을 조합 최적화 문제로 다룬다[3]. 메타 휴리스틱 알고리즘은 현실세계의 복잡한 문제를 간단히 해결하고자 한다. 최근에 메타 매칭을 사용한 온톨로지 정렬 알고리즘에서는 재현율과 정확률의 최대화를 목표로 제안하였다. 그러나 사용한 휴리스틱 알고리즘은 입력해야 하는 파라미터가 많기 때문에 계산이 복잡하며 특정 도메인의 다양한 데이터에 따른 조율작업이 필요하여 최적의 정렬 결과를 얻는 것이 어렵다.

이 논문에서는 메타 휴리스틱 알고리즘 DCS(Discrete Cuckoo Search)를 기반을 둔 온톨로지 정렬알고리즘을 제안한다. DCS는 이산적인 공간에서 조합 최적화 문제를 다루기 위하여 제안되었다. Levy Flight 분포를 따른 탐색을 통하여 효율적이고 조율하는 파라미터의 수가 적고 계산이 간단하며 이산적인 공간에서 기존 휴리스틱 알고리즘보다 좋은 결과를 보여주었다[4]. 이 논문에서는 군집화에 속한 각 개체를 후보 정렬로 표현하고 개체들은 간단하고 빠른 협력과 경쟁을 통해 정확률과 재현율을 최대화 하고자한다. 제안한 알고리즘의 성능을 보여주기 위하여 OAEI 2012에서 제공한 벤치마크 데이터를 사용하여 OAEI 2012에서 제안한 알고리즘들과 기존 메타 매칭을 사용한 온톨로지 정렬 알고리즘을 비교 평가한다.

이 논문은 다음과 같이 구성된다. 2장에서는 기존 온톨로지 정렬 시스템 및 메타 매칭에 대해서 설명한다. 3장에서는 온톨로지 및 온톨로지 정렬에 대하여 알아본다. 4장에서는 이 논문에서 제안한 이산 Cuckoo Search를 사용한 온톨로지 정렬 알고리즘을 소개한다. 5장에서는 제안한 알고리즘 실험 및 기존 시스템들과 비교평가하며 5장에서는 결론과 향후 연구를 서술한다.

2. 관련 연구

최근 온톨로지 이질성을 해결하기 위해 여러 가지 온톨로지 정렬 시스템이 개발되었다. 기존 단일 유사도 측정 시스템에서는 온톨로지 정보 표현의 모호성으로 사용자가 원하는 결과를 얻는 데 한계가 있다. 최근 온톨로지 정렬 시스템은 여러 가지 유사도 측정 기법을 결합한 복합시스템이 제안되었다. 서로 다른 정보 타입(label, text, description, structure, rules)을 처리하고 각 항목의 유사도 값을 최적의 가중치 조합에 기반하여 결과의 정확성을 높였다. 대표적인 시스템으로는 RiMOM[5], COMA[6], COMA++[7], QuickMig [8], FOAM[9], iMAP[10] 및 OntoBuilder[11] 등이 있다. 기존 복합시스템에서는 온톨로지 개념의 양에 따라 매트릭스를 생성하고 전수조사를 진행하여 온톨로지 정렬을 구축하기에 온톨로지 개념의 양의 규모가 커짐에 따라 계산이 복잡해지고 걸리는 시간이 기하급수적으로 증가되며 오류도 발생이 증가한다. 이러한 문제를 해결하기 위하여 최근에는 메타 매칭 기법을 사용하는 연구가 대두되고 있다.

메타 매칭은 메타 휴리스틱 알고리즘을 사용하여 온톨로지 정렬을 조합 최적화 문제로 다룬다. 메타 휴리스틱 알고리즘은 현실세계의 복잡한 문제를 간단히 해결하기 위하여 제안되었다. 메타 매칭을 이용하여 온톨로지 정렬을 해결하는 방법은 두 가지로 유형으로 나뉜다. 첫 번째 유형은 메타 휴리스틱 알고리즘을 사용하여 특정 도메인의 온톨로지 정렬에 사용되는 유사도 기법, 가중치 및 한계치 등 최적인 조합을 선택하는 방법이다. 대표적인 알고리즘은 Martinez-Gil이 제안한 GOAL이다[12]. 이는 유전 알고리즘을 사용하여 도메인에 따른 유사도 측정 기법들의 가중치 조합을 구하여 자동으로 처리하는 기법이다. Xingsi Xue는 GOAL을 확장하여 다중 목적 함수를 사용하는 MOEA/D를 제안하여 기존 단일 목적함수의 결과보다 좋은 성능을 보였다[5]. 두 번째 유형은 온톨로지 정렬을 직접 최적화 문제로 다루는 것이다. 이는 대량의 온톨로지 데이터에 대한 전수조사를 피할 수 있다. 또 특정 도메인의 다양한 데이터에 적합한 목적함수 및 유사도 측정 함수를 쉽게 수정 및 대체할 수 있다. 개체마다 독립적인 계산을 지원하여 병렬처리와 개체들 사이의 협력과 경쟁을 통해 전역적인 탐색이 가능하며 종료 조건을 지정하여 원하는 시각에 실행을 종료할 수 있다. 이러한 장점들로 인하여 Jurgen Bock은 DPSO (Discrete Particle Swarm Optimization) 기반을 둔 온톨로지 정렬 시스템을 제안하였다[13]. DPSO 알고리즘은 PSO를 변경한 알고리즘으로서 최적의 속성 집합 선택을 목표로 제안되었으며 비슷한 개념으로 인해 온톨로지 정렬의 대응관계를 찾는 작업에 사용되어 연구되었다. 그러나 DPSO는 사용되는 파라미터가 많기 때문에 계산이 복잡하고 특정 도메인 다양한 데이터에 따른 조율작업이 요구되며 온톨로지 정렬에 좋은 성능을 보여주지 못했다.

3. 온톨로지 및 온톨로지 정렬

3.1 온톨로지 및 온톨로지 정렬

온톨로지란 “공유하기 위한 개념들의 개념화를 형식적이고, 명백하게 설명해 놓은 명세서”이다[1]. 온톨로지는 특정된 도

메인의 개념들과 이들 개념 간의 상호관계를 식별하고 계층적으로 보여주며 수식 O=(C, P, I)로 표시한다. O는 온톨로지를 표시하고 C는 개념들의 집합이고, P는 속성들의 집합 즉 영역 개념 간의 관계들이며, I는 인스턴스들의 집합 즉 개념이 현실 생활에서 대응되는 대상들을 나타낸다. 이 논문에서는 개념, 속성, 인스턴스를 총칭하여 엔터티라고 부르며 수식(1)과 같이 표시한다. E는 온톨로지에 속하는 엔터티들의 집합이며 R는 엔터티들 간에 이루는 관계(=, ⊆, ⊥)들의 집합이다.

$$O=(E,R) \tag{1}$$

온톨로지 매칭은 두 온톨로지 간의 엔터티들의 유사도 값에 기반을 두어 대응관계를 찾아 연관관계를 맺는 작업이며 함수로 표시하면 수식(2)와 같다.

$$m=f(id,e_1,e_2,r,s)\ where\ e_1\in E_1,e_2\in E_2 \tag{2}$$

수식(2)에서 id은 두 온톨로지의 엔터티로 이룬 대응관계의 식별자고, e_1,e_2는 두 온톨로지 엔터티의 집합 E_1,E_2에서 임의로 선택된 엔터티들이다. r은 두 엔터티의 관계를 나타내며 일반적으로 엔터티들이 이루는 대응관계는 크게 등치(= equivalence), 포함(⊆ subsumption), 무연관(⊥ disjointness)로 분류된다. s는 두 엔터티의 관계를 나타내는 유사도 값을 보여준다. s는 [0, 1]의 범위를 가지며 0은 두 엔터티 간에 같지 않음을 나타내고 1은 두 엔터티가 등치관계를 나타낸다. 이 연구에서는 1:1 등치관계를 가지는 매칭을 고려한다.

온톨로지 정렬은 두 온톨로지 사이의 엔터티들이 관계를 이루는 대응관계들의 집합을 찾는 작업을 말하며 함수로 표시하면 수식(3)과 같다.

$$A'=f(O_1,O_2,A,p,r) \tag{3}$$

수식(3)에서 O_1,O_2는 정렬에 참여하는 두 온톨로지를 표시하고 A는 사전에 결정된 부분적인 정렬이다. p는 파라미터들의 집합이고, r은 외부인 자원을 가리키며 A'는 함수에 인해 반환된 새로운 후보 정렬을 표시한다. 후보 정렬의 평가 함수는 수식(4)와 같이 대응관계 유사도 값들의 총합이다.

$$F(A)=\sum_{m_i:A} f(m_i) \tag{4}$$

3.2 유사도 함수

이 논문에서는 문자열 매칭 기법인 레벤슈타인거리와 SMOA거리를 이용하여 유사도 값을 구한다. 레벤슈타인거리는 편집 거리 기법 중의 일종으로서 두 문자열지 간에 하나의 문자열에서 다른 문자열로 전환할 때 필요한 삭제, 추가, 변환의 최소 편집 횟수를 계산한다[14].

$$lev(s_1,s_2)=\frac{dist(s_1,s_2)}{\max(|s_1|,|s_2|)} \tag{5}$$

$|s_1|,|s_2|$는 문자열 s_1,s_2의 길이이며 $dist(s_1,s_2)$는 문자열 s_1,s_2간의 레벤슈타인 길이이다.

SMOA 거리는 문자열 편집 거리 유사도 측정 기술이며 수식(6)과 같다[15]. $Comm(s_1,s_2)$ 문자열 s_1,s_2 간에 같은 부분의 길이를 반환하고, $Diff(s_1,s_2)$는 문자열 s_1,s_2 간에 다른 부분의 길이를 반환하며 $WinklerImpr(s_1,s_2)$는 기존 $Winkler$함수를 확장하였다.

$$smoa(s_1,s_2)=Comm(s_1,s_2)-Diff(s_1,s_2)+WinklerImpr(s_1,s_2) \tag{6}$$

4. 이산 Cuckoo Search를 이용한 온톨로지 정렬

4.1 이산 Cuckoo Search

1) 둥지(nest)

둥지는 군집화에서 하나의 개체이다. 둥지의 개수 N는 군집화의 크기를 결정하며 CS에서 초기화 할 때 개수는 고정된다. 군집화에서 새로운 다른 둥지에 의해 대체되거나 버림받는다. 이 논문에서 둥지는 군집화에서 후보 정렬을 포함한 하나의 개체로 간주한다. 일반적으로 하나의 둥지에는 여러 개의 알이 놓여질 수 있지만 이 논문에서는 각 둥지마다 하나의 알만 포함된다고 가정한다.

2) 알(egg)

뻐꾸기는 하나의 둥지에 오직 하나의 알을 탁란할 수 있다고 가정하면 아래와 같다.

a) 각 알은 둥지 안에서 군집화에 속한 개체의 해로 표현한다.
b) 뻐꾸기 알은 군집화에서 새로운 후보 해로 표현되며 기존 둥지의 알을 대체하거나 주인 새에게 발견되었을 경우 버림받는다.

각 알은 하나의 후보 정렬을 나타내며 각 알마다 d차원을 가지는 벡터로서 크기는 (d∈(0, min(|E1|,|E2|))) 무작위로 결정된다. 각 차원은 수식(2)과 같이 두 온톨로지 간의 임의로 선택된 엔터티들로 이루어진 대응관계를 표시하며 이들로 이루어진 후보 정렬을 표시하면 수식(7)과 같다.

$$A_i=m_0,m_1,...,m_{d-1} \tag{7}$$

수식(7)에서 A_i는 i번째 둥지에 포함된 해, 즉 후보 정렬을 표현하며 m_k은 알의 k+1번째 차원 즉 대응관계를 말한다.

3) 목적함수

각 해는 탐색공간에서 목적 값을 수치적으로 표현하며 적합도는 목적함수에 결정된다. 이 논문에서는 온톨로지 정렬의 적합도와 크기를 고려하여 목적함수 수식(8)을 정의하여 후보정렬을 평가한다.

$$Fitness(A) = \lambda * (\min(|E_1|,|E_2|) - |A|) + (1-\lambda) * F(A) \quad (8)$$

수식(8)에서 $Fitness(A)$는 후보 정렬 A의 적합도 값을 반환한다. 수식(8)에서 파라미터 λ는 가중치로서 값이 클수록 후보정렬을 이루는 대응관계 개수에 집중하여 탐색하여 정확성이 낮아지며 반대로 가중치 값이 작아지면 정확성에 집중해서 탐색하고 대응관계 개수 크기가 낮아진다.

4) Levy Flights

Levy Flight은 장기적으로 좋은 해를 포함한 지역에 집중적으로 탐색하고 가끔 큰 스텝을 거쳐 검색되지 않은 지역을 탐색함으로써 강화 및 다양화 전략들의 균형을 조절하는 중요한 역할을 한다. 스텝의 크기는 성능에 영향을 미치며 검색하는 공간 범위 안에 포함되어야 한다. 개체는 Levy Flight에 의해 여러 개의 작은 스텝과 가끔 큰 스텝으로 인해 공간에서 이동한다.

이 논문에서는 Levy Flight 범위를 [0,1]로 정한다. 이 간격에서 적당한 스텝 길이를 선택하고 Levy Flight에 분포에 따른 비례에 따라 최적의 둥지 안에 알을 모방한다. 새로운 세대마다 뻐꾸기 알은 rand(pt*dbest,n)차원을 (pt는 Levy Flight([0,1]) 분포에 따른 확률) 가지고 최적의 둥지 안에 있는 알을 pt 비례만큼 모방하고(수식(9)) 나머지는 무작위로 만들어 다른 둥지에 탁란한다.

$$A_n = b_1, b_2, b_3, ..., m_{(n,k)}, m_{(n,k+1)} \quad (9)$$

4.2 DCS를 이용한 온톨로지 정렬

Fig. 1은 DCS를 사용한 온톨로지 정렬의 프로세스를 나타내며 크게 초기화 단계와 군집화 진화 단계 두 가지로 나눈다. 초기화 단계에서는 군집화의 개체에 대하여 정의하고 소개하며 군집화 진화 단계에서는 개체 진화 과정에 대한 계산 과정을 설명한다.

1) 초기화 단계

Table 1은 초기 모집단 과정을 보여준다. 이 과정에서 군집화 각 개체마다 무작위로 초기화 된다. 즉 각 개체의 차원마다 무작위로 두 온톨로지 간의 엔터티로 대응관계를 이루어 초기화를 마치고 수식(2)를 이용하여 유사도 값을 구하고 저장하고 차원의 수치에 따라 반복한다. 개체 즉 후보 정렬이 생성된 후 유사도 값에 의해 각 차원들을 오름차순으로 배열한 후 목적함수(수식(8))에 의해 후보 정렬의 적합도를 구하고 저장한다. 마지막으로 최적의 후보 정렬을 구하고 다음 세대로 넘어간다.

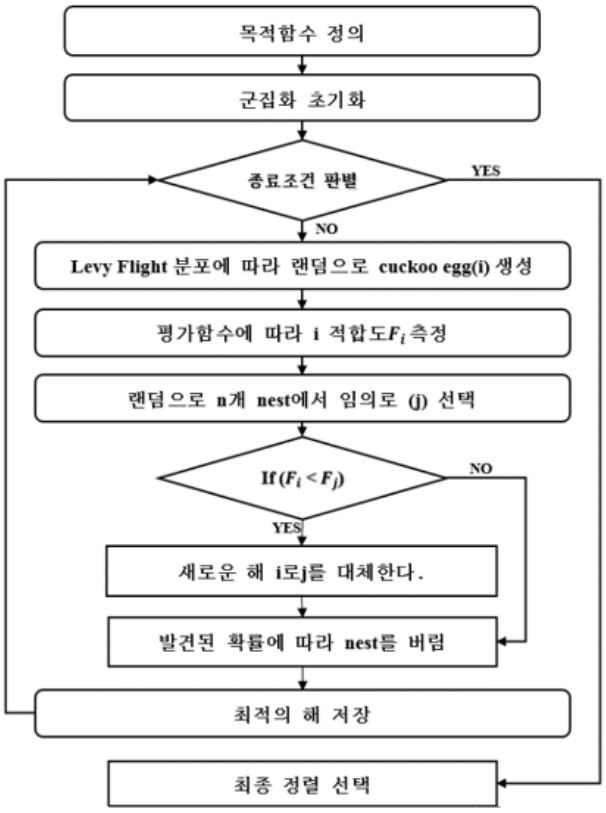

Fig. 1. The process of DCS

Table 1. The initialization of ontology alignment based on DCS

```
01: Input: N the number of Nests
02:       n = min(| E1 |, | E2 |)
03:       for i = 1 to N do
04:           d = rand(0,n)
05:           for j = 1 to d do
06:               /*randomly selects entities and evaluate
07:               similarity (have not already been selected)*/
08:               mi,j = f(idi,j,e1,e2,r,s)  where e1 ∈ E1 and e2 ∈ E2
09:               Ai = Ai ∪ mi,j
10:           end for
11:           order mj by the similarity s in Ai
12:           evaluate coordinate Alignment F(Ai)
13:           keep the best Nest
14:       end for
```

2) 군집화 진화 단계

Table 2는 I번의 횟수를 통해 진화를 진행한다. 새로운 세대마다 뻐꾸기는 Levy Flight 분포를 따르는 비율에 따라 최적의 알을 모방하고 탁란한다. 새로 생성된 뻐꾸기 알은 목적함수(수식(8))에 의해 평가되고, 최적의 해와 비교하였을 때 보다 나은 퀄러티를 보인다면 최적의 해를 대체한다. 뻐꾸기 알이 주인 새에게 (발견된 확률 pa) 발견되면 버림

받는다. 최적의 해를 담은 둥지를 찾고 저장하고 종료조건을 확인하고 계산을 반복한다.

Table 2. Ontology Alignment based on DCS

```
01: objective function f(x), x = (x1,...,xd)^T
02: //Generate Initialize population of n host Nests xi(i = 1,...,n)
03:while(t < tmax)
04:   for i = 1 to n do
05:       pi ⇐ levyflight(0,1);
06:       lbegin ⇐ pi ⊗ dbestnest;
07:       di ⇐ rand(lbegin ,n)
08:       for k = 1 to lgest do
09:           Xi ⇐ Xi ∪ mbest_k;
10:       end for
11:       for j = 1 to di − lbegin do
12:           /* cuckoo start from their nest search new nest */
13:           mi = get a new cuckoo randomly(j);
14:           Xi ⇐ Xi ∪ mi;
15:       end for
16:       Build F(Xi)    //Evaluate Xi fitness
17:       A bandon a fraction () of worse nests and built new ones;
18:       if(F(Xi) < F(Xbest))
19:          /* Replace best by the new solution */
20:          Keep the best solutions or nests with fitness solutions;
21:          Rank the Alignment by the similarity;
22:       end if
23:   end for
24:end while
25: post process results and visualization;
```

5. 실험 및 평가

5.1 실험환경 및 데이터

이 논문에서는 제안한 알고리즘의 성능을 평가하기 위하여 OAEI 2012에서 제공하는 벤치마크 데이터를 사용하였고 Alignment OAEI API를 이용하여 실험 평가하였다[17]. 벤치마크 데이터는 하나의 온톨로지를 110가지 형태로 만들어 제공한다. 100시리즈는 온톨로지의 개념정보는 유지하고 구조를 수정한 형태이고, 200시리즈는 온톨로지 구조정보를 유지하고 개념을 동의어나 외래어로 수정하였으며, 300시리즈는 온톨로지의 실제 상황에서 사용되는 형태로 제공된다. 구현 환경은 Windows 7 x86 운영체제 인텔 코어 i3 CPU와 4GB 메모리 및 JDK 7이다.

알고리즘의 결과 값이 효율적으로 설정될 수 있도록 모집단의 크기, 발견된 확률, 반복 횟수를 Table 3과 같이 적용한다.

Table 3. Initial values for parameter setting

파라미터	모집단 크기(N)	발견된 확률(pa)	반복 횟수(I)	가중치(λ)
초기값	100	50%	100	0.65

온톨로지 정렬 성능 평가 척도로는 정확률(Precision)과 재현율(Recall)을 이용한다. 정확률은 검색된 정렬의 총 항목 중에서 관련 정렬 항목에 일치하는 비율을 보여주고 재현율은 관련 정렬 총 항목 중에서 검색된 정렬의 항목이 일치하는 것을 비율로 보여준다.

$$Precision = \frac{\{Relevant\ Alignment\} \cap \{Rretried\ Alignment\}}{\{Retrieved\ Alignment\}} \quad (10)$$

$$Recall = \frac{\{Relevant\ Alignment\} \cap \{Rretried\ Alignment\}}{\{Relevant\ Alignment\}}$$

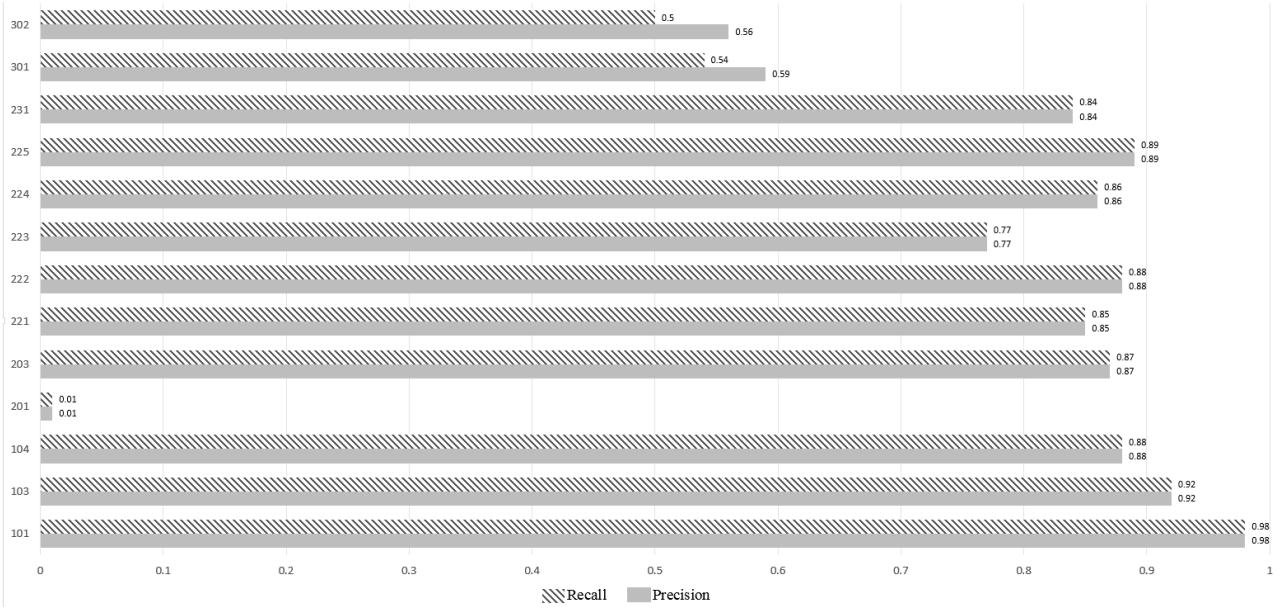

Fig. 2. The result of DCS in the OAEI 2012 benchmark data set

5.2 실험 및 검증

Fig. 2는 제안한 알고리즘의 사용가능성을 평가하기 위하여 단일 문자열 거리 측정 기법과 OAEI 2012 벤치마크 데이터를 이용하여 측정하였다. 100시리즈 온톨로지를 측정한 정확율과 재현율은 평균적으로 90% 이상의 좋은 성능을 보여준다. 101온톨로지의 의미정보는 유치하고 구조만을 바꾼 경우이기에 측정 기법에 적합하다. 200시리즈 데이터 측정 결과에서는 온톨로지 201을 제외하고 평균적으로 80%이상의 결과를 보여준다. 이는 기존 온톨로지 구조정보를 유지하고 개념을 유사어로 수정하였기 때문이다. 마지막으로 온톨로지의 실제 상황에서 사용되는 300시리즈 데이터 측정 결과는 50~60% 사이의 상대적으로 낮은 값을 보여준다. 이는 온톨로지의 모호성으로 단일 유사도 측정 기법으로 이용하여 만족한 결과를 얻지 못함을 보여준다. 그러나 제안한 알고리즘은 사용자가 상호 작용이 선택 가능한 유연성을 지원하여 특정 도메인 다양한 데이터에 대하여 합당한 매칭 방법을 적용하여 (WordNet, 택사노미 기법) 높은 성능 결과를 얻을 수 있다.

Fig. 3은 제안한 알고리즘과 OAEI 2012에서 참여한 정렬 시스템과 성능을 비교한 결과이다. OAEI 2012에서 AROMA가 가장 우수한 성능을 보여준다. 온톨로지 100시리즈에서 제안한 알고리즘은 AROMA, WeSeE 다음으로 높은 정확률을 보여준다. 온토롤지 200, 300시리즈에서는 개념을 유사어로 수정하였기에 기존 알고리즘보다 상대적으로 낮은 정확율을 보여준다. 온톨로지 100, 200시리즈에서는 30~40% 제고된 재현율을 보여주며 300시리즈에서는 AROMA와 MEDLEY 다음으로 높은 평균 이상의 재현율을 보여준다. 제안하는 알고리즘은 전체적인 성능을 비교하였을 때 좋은 성능을 보여준다.

Fig. 4는 제안한 DCS 알고리즘의 타당성을 평가하기 위하여 DPSO 알고리즘을 직접 구현하여 비교한 결과이다. 비교된 두 알고리즘은 같은 목적함수와 유사도 측정 기법을 적용하였다. 우선 온톨로지 300시리즈 데이터 셋에 대한 결과를 보았을 때 개념의 이질성과 모호성으로 인하여 단일 문자열 측정 기법을 사용하여 측정하였기에 좋은 결과를 보여주지 못하였다. 그러나 온톨로지 100시리즈, 200시리즈 데이터 셋에 대한 결과에서는 제안한 알고리즘은 기존 DPSO 알고리즘보다 30~50% 성능이 향상되었다. 부분적인 이유로는 DCS가 DPSO(β, ɣ, κ, σ 등)보다 조율해야 하는 매개변수가 적기 때문이다. DCS 에서는 기본적으로 pa 하나의 매개변수가 있음을 볼 수 있고 군집화 집중률에 영향을 미치지 않는다. 그러므로 일반적으로 pa를 조율할 필요성이 없어진다. 그러므로 제안된 알고리즘은 DPSO에 비해 조율해야 하는 파라미터의 개수가 적고 계산이 간단하여 실행 시간이 적으며 좋은 성능을 보여준다.

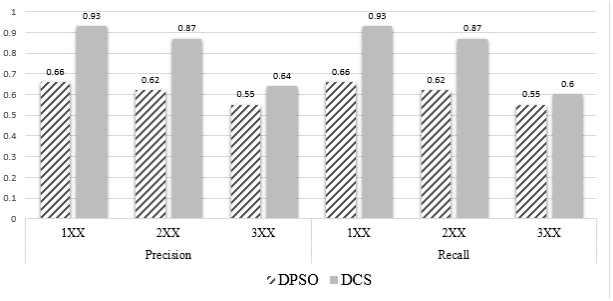

Fig. 4. The comparisons between the DCS and DPSO

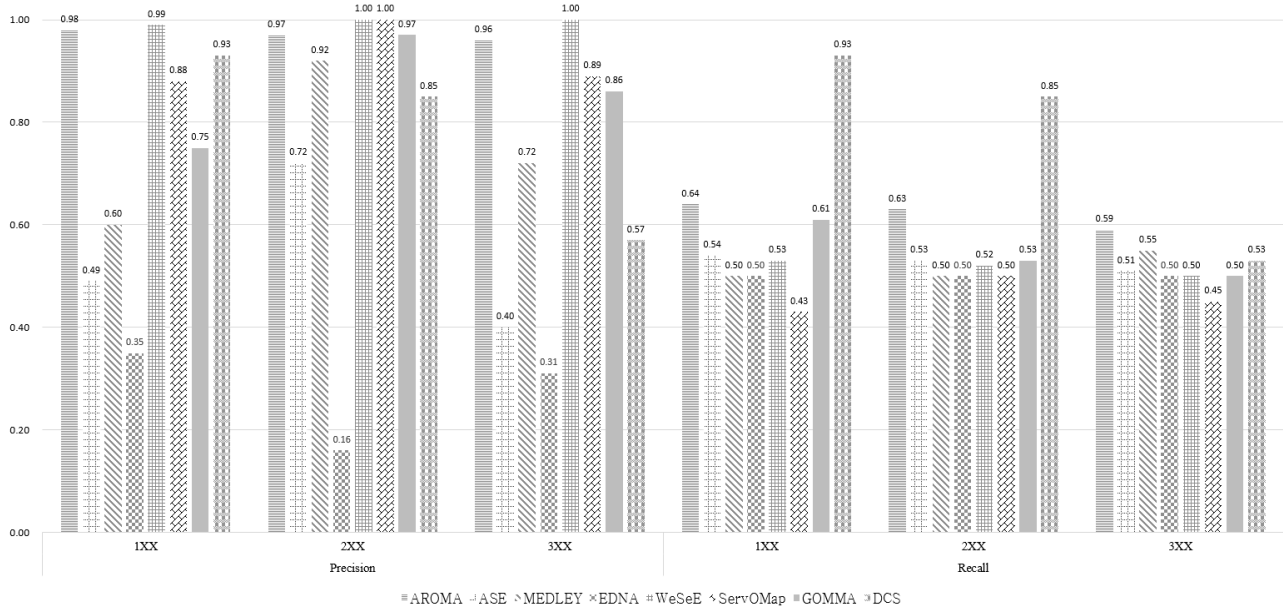

Fig. 3. The comparison between the proposed approach and the participants of OAEI 2012

6. 결론 및 향후 연구

온톨로지 정렬은 같은 도메인을 다루는 온톨로지 간의 상호운영성을 지원하기 위한 전제 조건이다. 온톨로지 정렬은 온톨로지 개념의 양이 많아지면서 계산이 복잡해지고 걸리는 시간이 기하급수적으로 증가한다. 이 논문에서는 DCS를 사용한 온톨로지 정렬 알고리즘을 제안한다. 이는 온톨로지 정렬의 특정 도메인 데이터의 상호작용 유연성 및 파라미터 조율의 간편성을 고려하고 정확률과 재현율의 최대화에 목적이 있다. 사용한 DCS에서 둥지에 포함된 알은 후보 정렬로서 Levy Flight에 분포를 통해 탐색한다. Levy Flight 알고리즘 모델은 단순하고 조율해야 하는 파라미터가 적으며 좋은 해를 포함한 지역에 집중적으로 탐색하고 가끔 큰 스텝을 거쳐 검색되지 않은 지역을 탐색함으로써 효율적인 강화(intensification) 및 다양화(diversification) 전략을 보여준다. 제안된 알고리즘의 성능을 평가하기 위하여 OAEI 2012에서 제안된 시스템과 비교하였으며 문자열 기반 기법을 사용하여 대상이 되는 100, 200시리즈 온톨로지에는 높은 성능을 보였다. 이는 제안한 알고리즘이 대상되는 데이터에 효율적인 검색을 보여준다. 기존 메타 휴리스틱 알고리즘 DPSO 기반 온톨로지 정렬 알고리즘과 비교한 결과 제안한 알고리즘이 40% 향상된 성능을 보여 주었다. 이로부터 제안한 알고리즘이 사용된 파라미터가 적고 계산이 간단하여 걸리는 실행 시간이 짧으면서도 효율적인 탐색을 보여준다.

향후 연구로 제안한 알고리즘을 확장하여 특정 도메인 데이터 온톨로지 정렬에 합당한 목적함수를 연구하고 온톨로지 정렬 멀티 목적함수를 사용하며 자동적인 파라미터 배치를 적용하여 좋은 성능을 보여주고자 한다.

References

[1] Gruber, Thomas R., “A translation approach to portable ontology specifications,” Knowledge acquisition Vol.5, No.2, pp.199-220, 1993.

[2] Granitzer, Michael, et al., “Ontology alignment-a survey with focus on visually supported semi-automatic techniques,” Future Internet Vol.2, No.3, pp.238-258, 2010.

[3] Xue, Xingsi, Yuping Wang, and Weichen Hao, “Using MOEA/D for optimizing ontology alignments,” Soft Computing, pp.1-13, 2013.

[4] Ouaarab, Aziz, Belaïd Ahiod, and Xin-She Yang, “Improved and Discrete Cuckoo Search for Solving the Travelling Salesman Problem,” Cuckoo Search and Firefly Algorithm. Springer International Publishing, pp.63-84, 2014.

[5] Li, Juanzi, et al., “Rimom: A dynamic multistrategy ontology alignment framework,” Knowledge and Data Engineering, IEEE Transactions on Vol.21, No.8: pp.1218-1232, 2009.

[6] Do, Hong-Hai, and Erhard Rahm, “COMA: a system for flexible combination of schema matching approaches,” Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002.

[7] Aumueller, David, et al., “Schema and ontology matching with COMA++,” Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005.

[8] Drumm, Christian, et al., “Quickmig: automatic schema matching for data migration projects,” Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007.

[9] Ehrig, Marc, and York Sure, “Foam-framework for ontology alignment and mapping results of the ontology alignment evaluation initiative,” Integrating Ontologies Workshop Proceedings, Vol.72, 2005.

[10] Dhamankar, Robin, et al., “iMAP: discovering complex semantic matches between database schemas,” Proceedings of the 2004 ACM SIGMOD international conference on Management of data. ACM, 2004.

[11] Gal, Avigdor, Giovanni Modica, and Hasan Jamil, “Ontobuilder: Fully automatic extraction and consolidation of ontologies from web sources,” Data Engineering, 2004. Proceedings. 20th International Conference on. IEEE, 2004.

[12] Martinez-Gil, Jorge, Enrique Alba, and José F. Aldana-Montes, “Optimizing ontology alignments by using genetic algorithms,” Proceedings of the workshop on nature based reasoning for the semantic Web. Karlsruhe, Germany, 2008.

[13] Bock, Jürgen, and Jan Hettenhausen, “Ontology Alignment using Discrete Particle Swarm Optimisation.”

[14] Levenshtein, Vladimir I., “Binary codes capable of correcting deletions, insertions and reversals.” Soviet physics doklady. Vol.10, 1966.

[15] Stoilos, Giorgos, Giorgos Stamou, and Stefanos Kollias., “A string metric for ontology alignment,” The Semantic Web-ISWC 2005. Springer Berlin Heidelberg, pp.624-637, 2005.

[16] Yang, Xin-She, and Suash Deb., “Cuckoo search via Lévy flights,” Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, 2009.

[17] Ontology Alignment Evaluation Initiative, (2012), <http://oaei. ontologymatching.org>.

한 군

e-mail : jeremie_han@korea.ac.kr

2006년 연변과학기술대학교 컴퓨터학과(학사)

2013년~현 재 고려대학교 컴퓨터·전파통신공학과 석사과정

관심분야 : 시맨틱 웹, 온톨로지, 소프트웨어 공학, 휴리스틱 알고리즘

정 현 준

e-mail : darkspen@korea.ac.kr
2008년 삼육대학교 컴퓨터과학과(학사)
2010년 숭실대학교 컴퓨터학과(석사)
2011년~현 재 고려대학교 컴퓨터·전파통신공학과 박사과정
관심분야 : 시맨틱 웹, 온톨로지, 메타데이터, 레지스트리, 소프트웨어 공학, 센서 네트워크

백 두 권

e-mail : baikdk@korea.ac.kr
1974년 고려대학교 수학과(학사)
1977년 고려대학교 산업공학과(석사)
1985년 Wayne State Univ. 전산학과(석사)
1989년~2007년 한국정보과학회(이사/평의원/부회장)
1986년~현 재 고려대학교 컴퓨터·전파통신공학과 교수
1991년~현 재 한국시뮬레이션학회(이사/부회장/감사/회장/고문)
1991년~현 재 ISO/IEC JTC1/SC32 전문위원회 위원장
2001년~현 재 도산아카데미 원장
2002년~2004년 고려대학교 정보통신대학 초대학장
2004년~2005년 한국정보처리학회 부회장
관심분야 : 데이터 모델링, 시뮬레이션, 데이터 공학, 소프트웨어 공학, 프로젝트 관리