

유전자알고리즘과 신경망을 이용한 웹 서비스 정보 클러스터링

김진성⁰ 정혜진 유춘식 김용성
전북대학교 공과대학 전자정보공학부
{kpiiju⁰, araves, csyoo, yskim}@chonbuk.ac.kr

Web Service Information Clustering using Genetic Algorithm and Neural Network

Jin Sung Kim⁰ Hye Jin Jeong, Chun Sik Yoo, Yong Sung Kim.
Division of Electronics and Information Engineering, Chonbuk National University

요 약

오늘날 웹서비스에 대한 정보 검색 시스템들은 UDDI 레지스트리에 대한 전문적인 지식을 필요로 한다. 즉 웹서비스를 명세하기 위해 사용된 카테고리 및 이에 대한 값, 이름 등을 사전에 알고 있어야만 검색이 가능하다. 그러나 일반 사용자들은 이러한 사전지식을 충분히 알고 있지 못하기 때문에 웹서비스에 대한 정보 검색이 쉽지가 않다. 그러므로 일반 사용자들을 위해 웹서비스에 대한 정보를 카테고리에 맞게 분류하면 웹서비스 검색을 보다 용이하게 할 수가 있다. 따라서 본 논문에서는 유전자 알고리즘과 신경망을 이용하여 보다 효율적으로 웹서비스 정보를 분류하는 클러스터링 기법을 제안한다.

핵심어 : 웹서비스, UDDI, 클러스터링, 유전자알고리즘, 신경망,

1. 서 론

최근 인터넷 관련 기술과 컴퓨터 관련 기술의 발달로 네트워크상에 분산되어 있는 자원과 정보를 보다 효과적으로 활용하기 위한 방안으로 분산 시스템에 대한 연구가 활발히 진행되고 있다. 이러한 분산 환경에서의 분산응용 개발을 지원하기 위해, 최근 W3C에서는 분산 환경에서의 표준 프로그래밍 모델로 웹 서비스(Web Service)라는 개념을 제시하였다[1].

현재 웹 서비스를 발견하기 위해서는 UDDI 레지스트리에 대한 전문적인 지식이 필요하다. 즉 웹 서비스를 명세하기 위해 사용된 카테고리(비즈니스 카테고리, 서비스 카테고리, tModels 카테고리)와 이에 대한 값, 이름(비즈니스 이름, 서비스 이름, tModels 이름)을 사전에 알고 있어야 한다. 그러나 일반 사용자는 이러한 사전 지식을 충분히 알고 있지 못하므로 키워드를 이용하는 일반적인 정보검색 기법처럼 웹 서비스에 대한 키워드 기반의 검색기법이 필요하다. 즉 서비스 요청자가 필요로 하는 최적의 웹 서비스를 검색하기 위해서 웹 서비스를 명세하기 위해 사용된 카테고리 및 이름, 웹 서비스 명세에 사용된 키워드 등을 함께 사용하는 새로운 방법론이 필요하다. 키워드를 이용한 UDDI 레지스트리 검색 기법에 대한 연구는 국내외적으로 전무한 실정이다.

본 논문의 목적은 사용자가 효과적으로 웹 서비스

정보를 검색할 수 있도록 웹서비스에 대한 정보들을 유전자알고리즘과 신경망을 이용하여 클러스터링 하는데 있다.

2. 관련연구

인공생명 개념을 이용한 키워드 기반의 지능형 UDDI 레지스트리 검색기법은 웹 서비스 발견을 위한 새로운 방법론이며, 분산응용 개발의 필요성이 증대되고 있는 현대사회에서 매우 중요한 핵심적 기술 중의 하나이다.

UDDI 레지스트리 검색에 관한 관련연구는 [2]가 있으며, 유전자알고리즘을 이용한 정보검색과 클러스터링에 관한 연구는 [3],[4],[5]가 있다. [2]는 XML을 기반으로 진보한 UDDI B2B 통합을 위한 방법을 연구하였다. 그리고 [3]은 집단의 크기를 크게 유지하면서 적합도 평가 과정을 줄이는 방안으로 클러스터링에 기반 한 효율적인 유전자 알고리즘을 제시하였다. [4]는 클러스터링 문제를 해결하기 위해서 유전자알고리즘을 사용하여, SICM, STCM, CSPM 3가지 모델을 사용하여 기술하였다. 특히 이 논문에서는 실험하고자 하는 p-Median 문제를 풀기 위하여 CSPM을 적용하였다. 한편 [5]는 신호 패턴에서 특징을 추출하여 유전자알고리즘과 신경망을 이용하여 신호를 분류하는 분류자를 구축하는 방법을 제시하였다. [6]은

동적인 그룹 사용자들에게 적합한 공명이론을 표준으로 한 신경망을 웹에 접근하는 기초로 하였다.

따라서 본 연구에서는 효과적으로 정보를 분류하기 위해서 유전자 알고리즘과 신경망을 이용하여 웹서비스의 정보에 대한 지속적인 학습을 통하여 웹서비스 정보를 클러스터링하고 사용자의 요구에 적합한 웹서비스 정보를 제공하고자 한다.

3. UDDI 레지스트리, 유전자 알고리즘, 신경망

본 논문에서 활용되는 데이터 구조인 UDDI 레지스트리 구조와 클러스터링에 적용하는 유전자알고리즘과 신경망에 관해서 알아본다.

3.1 UDDI 레지스트리 구조

UDDI의 데이터 구조는 <그림 1>과 같이 비즈니스 정보(Business entity), 비즈니스 서비스 정보, 바인딩 정보(Binding entity), tModel로 구성된다.

비즈니스 정보는 레지스트리에 등록하고자 하는 비즈니스 개체에 대한 서비스 제공자 정보이고, 비즈니스 서비스 정보는 등록되는 서비스에 대한 논리적인 정보를 의미한다. 또한 바인딩 정보는 웹 서비스에 접근하기 위한 기술적 정보를 표현하고, tModel은 비즈니스 서비스를 이용하기 위해 필요한 기술적인 정보나 웹 서비스의 분류/식별체계에 대해서 표현한다.

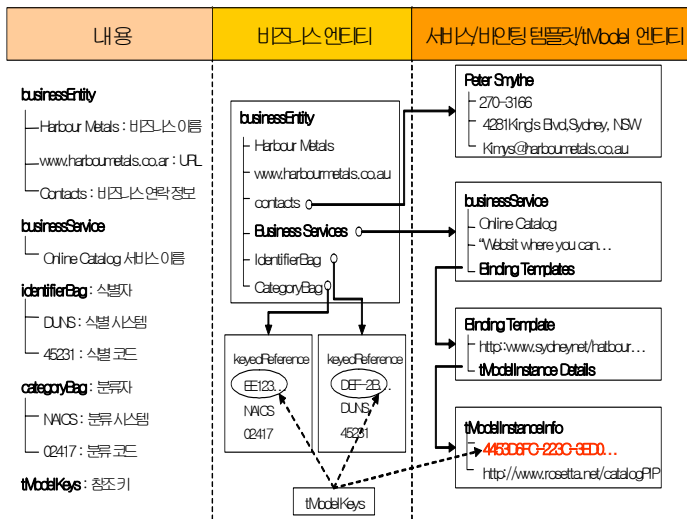


그림 1 UDDI의 데이터 구조

3.2 유전자알고리즘과 신경망

유전자 알고리즘(Genetic Algorithms)은 적응탐색(adaptive search) 방법론에 속하는 일반화된 발견적 해법이다.[7] 이 알고리즘은 자연계의 진화 과정에 기반을 두고 있으며, 환경에 잘 적응하는 유전자만을 선택하여 교배, 돌연변이에 의해 다음 세대의 개체군

을 형성하고 진화가 거듭될수록 그 환경에 더 적합한 유전자만이 남는 기본 개념을 가지고 있다. 유전자알고리즘의 기본 개념도는 <그림 2>와 같다.

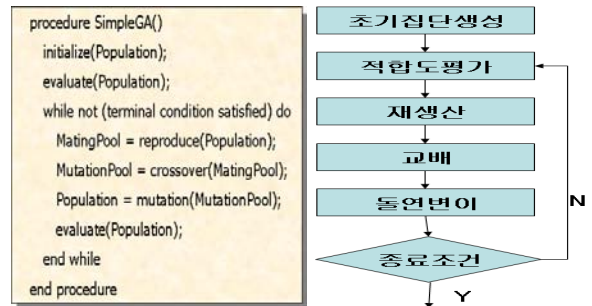


그림 2 유전자 알고리즘의 기본적인 흐름도

한편 1990년대 중반부터 마이크로파 영역에 적용되기 시작한 신경망 모델들은 학습과 최적화와 같은 경험적 지식에 입각한 작업들에 대한 알고리즘으로, 뇌의 본질에 관한 연구로부터 유도된 개념들에 기초한다[1]. 각 분야에 여러 가지 연구 데이터를 다시 활용하여 빠른 설계와 인식이 가능하다는 이유로 특히 패턴 인식 등에 많이 사용되고 있다. 신경망의 구조는 다음 <그림 3>과 같다.

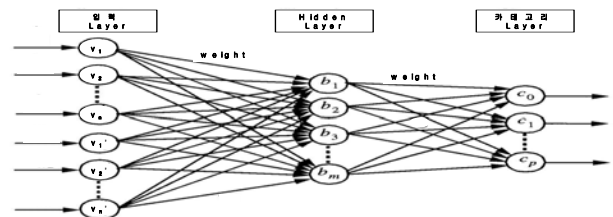


그림 3 신경망의 구조

4. 웹 서비스 정보 클러스터링

본 논문에서는 UDDI 레지스트리에 저장된 각 비즈니스 엔티티와 웹서비스에 대한 정보들을 추출하고, 이를 유전자알고리즘과 신경망을 이용하여 클러스터링하는 기법을 제안한다.

4.1 유전자알고리즘과 신경망을 이용한 클러스터링 절차

UDDI 레지스트리의 정보를 유전자알고리즘과 신경망을 이용하여 클러스터링하는 절차는 다음과 같다.

- ① UDDI 레지스트리에서 정보를 추출한다.
- ② 유전자알고리즘을 이용하여 유사 정보 패턴을 찾는다.
- ③ 신경망을 이용하여 유사규칙의 cutoff와 적합도를 평가한다.
- ④ 새롭게 생성된 규칙을 노드를 의사결정 트리의 해당 위치의 단말노드에 추가한다. 이를 유사 정보 패턴에 대한 마스크로 활용하여 Fail과 Pass 서브셋으로 정보가

분류된다.

<그림 4>는 유전자알고리즘과 신경망을 이용한 클러스터링 과정을 블록 다이어그램으로 도시한 것이다.

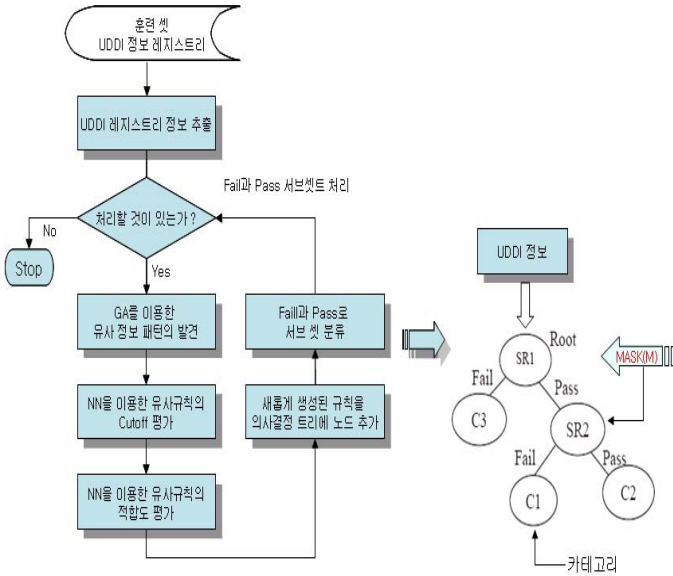


그림 4 GA/NN을 이용한 클러스터링 과정

따라서 본 논문에서는 UDDI 레지스트리의 정보를 대상으로 특징 정보를 추출하고 유전자 알고리즘으로 유사 정보 패턴을 발견한다. 또한 신경망을 이용하여 최종적으로 유사 규칙의 적합도를 평가하여 의사 결정 트리를 생성하고 이를 이용하여 레지스트리 정보를 클러스터링하는 방식을 제안한다.

4.2 UDDI 특징 정보 추출

UDDI 정보를 대상으로 웹 서비스별로 클러스터를 구성하기 위해서는 중요 특징 정보를 추출해야 한다. 이러한 특징 정보는 클러스터를 생성하기 위해서 UDDI 정보를 유일하게 구분할 수 있는 최소한의 특징으로 구성된다.

주요 특징 정보는 UDDI 정보에서 비즈니스 정보, 제품 정보, 서비스 정보, 기술정보 등에 대한 명사를 추출하여 구조화하고 불용어를 제거한 내용으로 구성한다.

4.3 퍼지 술어와 유사도

퍼지 술어로 UDDI의 특징 정보를 표현하고, 서로 다른 UDDI 레지스트리 정보 간에 유사성을 비교하기 위해서 유클리드 거리 유사도를 이용한다.

4.3.1 퍼지 술어

패턴정보 x에서 UDDI의 특징 정보 f_i 는 퍼지 술어 P_i 로 표현할 수 있다.

$$T(P_i) = \mu_1(x), \dots \dots \dots (식 4.1)$$

$$\mu_1(x) = \max \left\{ \frac{Area(f_i \cap f_j)}{Area(f_i) + Area(f_j) - Area(f_i \cap f_j)} \right\}$$

여기서 $\mu_1(x)$ 는 퍼지 술어 P_i 의 소속도이다.

단, 특징 정보 f_i 는 패턴 x의 특징 정보이고 μ_1 퍼지 값을 최대로 한다. 위 멤버십 함수는 직관적으로 μ_1 는 f_i 와 f_j 의 공통된 부분을 합으로 나눈 멤버십 함수임을 알 수 있다.

4.3.2 퍼지 유사도

서로 다른 UDDI 레지스트리 정보의 유사성을 비교하기 위해서 특징 정보의 집합을 $F = \{f_1, \dots, f_n\}$ 으로 표현하고, 하나의 UDDI 정보에서 특징의 집합과 일치하는 퍼지의 집합은 $\mu(w) = \{\mu_1(w), \dots, \mu_n(w)\}$ 으로 표현할 때 UDDI 정보 x와 UDDI 정보 y의 유클리드 거리 유사도는 다음과 같이 정의한다.

$$SD(x, y) = \sqrt{((\mu_1(x) - \mu_1(y))^2 + \dots + (\mu_n(x) - \mu_n(y))^2)} \dots \dots \dots (식 4.2)$$

4.4 의사결정 트리의 유사성 규칙

퍼지 유사도를 이용하여 의사 결정 분류 트리 노드의 유사성 규칙은 다음과 같이 표현된다.

```
IF (SD(x, m) > t) THEN visit the Pass - Child node
ELSE visit Fail-Child node,
```

여기서 $SD(x, m)$ 은 비교할 UDDI 정보 x와 패턴 m과의 유클리드 거리 유사성 정도를 의미한다. m은 특징 정보의 마스크 집합을 의미하고 t는 유사성 절단 값(alpha-cut)이다.

4.5 유전자 알고리즘을 이용한 유사규칙 발견

의사결정 분류 트리 안에 있는 각 노드들에 대한 최적의 유사 규칙을 발견하기 위한 유전자 알고리즘은 다음과 같다.

- 유전자 알고리즘의 인코딩 데이터는 UDDI의 비즈니스 정보, 제품정보, 서비스 정보의 속성에 의해서 표현되어지는 규칙에 대한 각 특징 정보 값을 사용한다.
- 초기집단은 의사결정 트리의 노드와 관련되어진 UDDI 정보 패턴 집합으로 부터 m개를 선택하여 학습 집합을 생성한다.
- 적합도 함수는 클러스터링의 성능을 결정하는 중요한 요소로 신경망을 이용한 유사 규칙의 적합도를 평가한다.
- 선택방법은 룰렛 휠(Roulette wheel)과 순위(Ranking) 선택방법을 사용한다.
- 교배방법은 라마키안(Lamarckian) 개념을 이용하여 의사결정 트리의 자식 노드 위치를 제어한다.

- 돌연변이는 non-uniform 돌연변이를 사용한다.
- 종결조건은 정해진 세대수가 생성되거나 정확한 수준의 적합도 규칙을 만족하면 종결조건으로 한다.

4.6 신경망을 이용한 적합성 평가

본 절에서는 신경망을 이용한 유사성 판단과 적합도 평가에 대한 방법을 기술한다.

4.6.1 유사성 판단과 적합도 규칙의 평가 알고리즘

유사성 판단값을 추출하는 방법과 적합성 규칙을 계산하는 알고리즘은 다음과 같다.

- 규칙 셋 안에 있는 모든 UDDI 정보를 대상으로 유클리트 거리 유사도 값을 계산한다.
- 가장 넓은 유사 분포에 의해서 유사도 값 사이의 갭(Gap)들을 정렬한다.
- 갭의 중앙값의 적합도를 P-mean으로 평가한다.
- 유사 규칙 cutoff의 가장 높은 적합도 값을 가진 중앙값을 선택하여 클러스터 센터를 생성한다.

4.6.2 적합도 평가

유사 규칙의 적합도 평가를 위한 계산식은 다음과 같은 방법으로 제안한다.

$$f = c_1g + c_2r - c_3(s_1 + s_2), \dots\dots\dots(식 4.3)$$

여기서 g는 유사 규칙 cutoff의 측면에서 두 클러스터 사이의 갭에 대한 크기, r은 패턴 셋의 크기 대 가장 작은 클러스터 크기에 대한 비율, s₁과 s₂는 두 클러스터의 확산 폭, 그리고 c₁, c₂, c₃는 상수 값을 의미한다.

4.7 학습 규칙을 적용한 의사결정 트리의 구축

본 논문에서 유전자 알고리즘과 신경망을 이용한 웹 서비스 정보를 대상으로 클러스터링을 하기 위해서 앞의 4.1에서 4.6까지 제시된 방법을 대상으로 학습 규칙을 개괄적으로 정의 하면 다음과 같다.

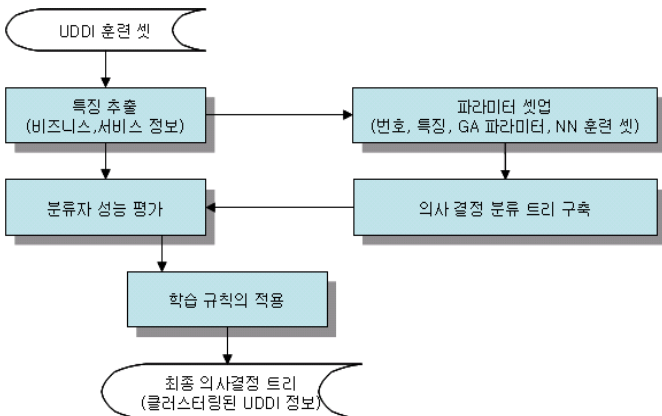


그림 5. 학습 규칙을 적용한 의사결정 트리 구축과정

5. 결론 및 향후 연구과제

본 논문에서 연구된 내용은 다음과 같이 나누어진

다.

첫째, 유전자 알고리즘을 이용한 웹서비스 UDDI 레지스트리의 특징 정보를 추출하고 유사패턴을 발견하기 위한 기법, 둘째, 추출된 검색체를 신경망에 적용하여 유사규칙 Cutoff의 평가와 유사규칙의 적합도 평가로 웹서비스를 클러스터링하는 방법을 제안하였다.

클러스터링할 때 유전자 알고리즘과 신경망을 이용함으로써 사용자가 웹 서비스 정보를 카테고리별로 효율적인 검색을 할 수 있을 뿐만 아니라, 더욱 다양하고 많은 자료를 검색 할 수 있을 것이다.

또한 유전자알고리즘과 신경망을 이용하여 UDDI 레지스트의 정보를 내용별로 클러스터링하고, 이를 이용한 로컬 DB를 구축함으로써, UDDI 정보를 검색하고자 하는 사용자들에게 편의를 제공하고 웹서비스의 활용을 활성화하여 기업 간의 활발한 전자상거래를 도모할 수 있을 것이다.

향후 본 논문에서 제안하고 있는 유전자알고리즘과 신경망을 이용하여 웹 서비스 정보를 클러스터링하는 알고리즘의 상세화와 이에 대한 구현, 실험/평가를 통한 효율성 입증을 위한 연구를 계속 수행할 것이다.

참고문헌

[1] Ying Huang, Jen-Yao Chung, "A Web services-based, framework for business integration solutions," Electronic Commerce Research and Applications, Vol.2, 2003.

[2] LIANG-JIE ZHANG, TIAN CHAO, JEN-YAO CHUNG, "XML-Based Advanced UDDI Search Mechanism for B2B Integration," Electronic Commerce Research, Vol.3, pp.25-42, 2003.

[3] 원흥희, 조성배, "클러스터링 기반의 효율적 유전자 알고리즘의 체계적인 성능 평가", 정보과학회 학술 발표논문집, 제29권, 제1호, pp.298-300, 2002.

[4] Yu-Chiun Chiou, Lawrence W.LAN, "Genetic clustering algorithms," European Journal of Operational Research, Vol.135, pp.413-427, 2001.

[5] Roshdy S. Youssif, Carla N. Purdy, "Combining genetic algorithms and neural networks to build a signal pattern classifier," Neurocomputing, Vol 61, pp 39-56, 2004.

[6] Santosh K, "Adaptive Neural Network Clustering of Web Users," IEEE, Vol.4, pp.34-40,2004.

[7] A. Patnaik , R.K. Mishra , "ANN Techniques in Microwave Engineering," IEEE, Microwave Magazine, vol 1, March 2000.