
진화연산을 통해 만들어지는 토픽맵

Evolutionary Topic Maps

김주호, Juho Kim^{*1}, 홍원욱, Won-Wook Hong^{**}, Robert Ian McKay^{***}

요약 진화연산은 최적화와 기계학습에 널리 사용되지만 또한 창조적이고 새로운 것을 만드는 데에도 사용될 수 있다. 본 논문에서는 지식을 표현하는 유연한 구조인 토픽맵에 주목하여, 새롭고 창의적인 토픽맵을 생성하는 토픽맵의 진화 시스템을 제안한다. 여기서는 만들어진 토픽맵이 유효한지에 대한 사람의 평가를 활용하는 대화형 진화 연산 방법(Interactive Evolutionary Computation)이 사용된다. 본 진화하는 토픽맵 시스템은 창의성을 도모하는 도구로서, 사용자들에게 새롭고 창의적인 지식을 떠올릴 수 있도록 도울 수 있을 것이다. 앞으로는 이 시스템에 보다 토픽맵에 정교한 사용자 인터페이스와 시각화 방법을 도입하고 기계학습을 활용하여 시스템의 진화 중에 나타나는 사용자의 피로를 크게 줄이는 방법을 연구할 것이다.

Abstract Evolutionary Computation is not only widely used in optimization and machine learning, but also being applied in creating novel structures and entities. This paper proposes evolutionary topic maps that can suggest new and creative knowledge not easily producible by humans. Interactive evolutionary computation method is applied into topic maps in order to accept human evaluation on feasibility of intermediate topic maps. Evolutionary topic maps are creativity support tools, helping users to encounter new and creative knowledge. Further work can greatly improve the system by providing more operations, preventing over-convergence, and overcoming user fatigue problem by providing more intuitive user interface, better visualization, and interpolation mechanisms.

This work was supported by the Brain Korea 21 Project and by Korea Research Foundation Grant Number (2008-KRF-331-D00943)

*주저자 : Stanford University Computer Science Department e-mail: juhokim@stanford.edu

**공동저자 : 서울대학교 컴퓨터공학부 e-mail: jradoo@gmail.com

***교신저자 : 서울대학교 컴퓨터공학부 교수; e-mail: rimsnucse@gmail.com

핵심어: 대화형 진화연산(Interactive Evolutionary Computation), 토픽맵(Topic maps), 유전 알고리즘(Genetic Algorithms), 지식표현(Knowledge Representation), 창의성 보조 도구(Creativity Support Tools)

1. 서론

인류의 학문 세계가 갈수록 복잡해지고 많아짐에 따라, 지식을 나타내고 시각화하는 문제가 중요한 화두가 되고 있다. 지식을 표현하는 방법의 하나로서 토픽맵은 표준화되어 있고, 매우 유연한 구조를 가지고 있다. 다만 토픽맵이 비록 지식을 표현하고 훑을 수 있는 좋은 방법이지만, 토픽맵을 만들고 관리하는 것은 쉽지 않은 작업이다. 우리가 토픽맵들간의 유사성과 관련성을 포착할 수 있다면 기존의 작업물로부터 새롭고 창의적인 토픽맵을 만들어낼 수 있을 것이다.

진화 연산(Evolutionary Computation)은 생물학적 진화에서 영감을 받아 자연 선택과 진화와 같은 개념들을 도입하고 있다. 진화 연산 기술은 최적화 분야에서 널리 사용되고 있고, 진화 과정에 생겨나는 다양성에 주목하여 창의적인 무언가를 컴퓨터가 만들어내도록 하는 시도도 이루어지고 있다. 본 논문에서는 진화 연산(EC)을 통해 컴퓨터로 하여금 새롭고 창의적인 토픽맵을 만들어내는 방법을 제안한다. 여기서 제안하는 구체적인 방법은 대화형 진화 연산(IEC)에 해당하는 것으로서, 미리 정의된 적합도 함수 대신 사람의 평가를 기준으로 적합도를 평가하도록 하는 것이다.

본 논문에서는 지식을 표현하는 방법인 토픽 맵을 진화 연산에 적용시켜봄으로써 새로운 창의성 도구의 가능성을 확인해보고 앞으로 가능한 논의에 대해서도 언급해본다.

2. 배경 이론

2.1 토픽맵(Topic Maps)

지식을 나타내는 여러 가지 방법들 중에 '맵'은 매우 효과적인 방법이다. 개념맵(concept map), 마인드맵(mind map), 토픽맵(topic map)이 널리 쓰이고 있다. 본 논문에서는 지식을 표현하고 그 관계를 나타내는 데에 주로 사용되는 토픽맵에 관해 논한다. 토픽맵은 질의를 처리하는데 용의하면서도, 특정한 의도 없이 자유롭게 돌아다닐 수도 있다. [3]. 이러한 점은 토픽맵의 구조상의 유연성과 확장성에서 비롯된다.

토픽맵은 지식을 표현하고 지식간의 관계를 나타낼 수 있게 해주고, 이미 존재하는 정보 원천과 바로 연결 지을 수 있게 되어 있다[2]. 토픽맵은 매우 복잡한 구조를 나타낼 수 있지만 기본적으로는 주제(topic), 관계(association), 그리고 발현(occurrence)의 간단한 개념으로 이루어져 있다[4]. 주제는 사람, 물건, 개념, 존재

등 기본적으로 어떤 것이든 지칭할 수 있다. 이런 토픽들은 서로 관계를 이룬다. 발현은 각 주제들이 나타나는 정보 자원들을 나타낸다. 토픽맵은 이러한 구조를 통해 데이터와 정보들간의 의미를 부여한다. 토픽맵은 또한 표준화된 맵 형식으로서의 위상을 가지고 있다.

2.2 진화 시스템(Evolutionary Systems)

진화연산(Evolutionary Computation)은 주어진 문제를 컴퓨터 프로그램의 형식으로 해답을 구하는 한 방법으로서 다윈의 자연선택설에서 영감을 받았다. 예를 들어 유전 알고리즘은 찾고자 하는 해를 다차원의 값으로 표현하여, 여기에 무작위적인 변형 혹은 다른 개체와의 교배를 가한다. 현 세대의 개체들이 다음 세대로 넘어갈 때, 좋은 해를 나타내는 개체는 살아남고 그렇지 않은 해는 버려지게 된다.

대체로 진화 알고리즘의 주안점은 최적화와 학습(machine learning)에 있다. 최적의 해를 찾거나 최소의 예측 오류를 가진 모델을 만드는 것이다. 그러나 진화 알고리즘은 최적의 해를 찾는 것에만 쓰일 수 있는 것이 아니라, 전혀 예측하지 않았던 새로운 해를 찾는 데에도 용이하다. 이 점 때문에 진화 알고리즘은 일종의 창의성 보조 도구로서의 가능성을 가지고 있다고 하겠다.

2.3 대화형 진화연산(Interactive Evolutionary Computation)

대화형 진화연산(IEC)에서는 적합도를 측정하는 데에 미리 정의된 함수를 사용하지 않고, 사람과의 상호작용을 통해 적합도를 결정한다. 이 방법은 사람의 선호도나 감정을 중심으로 적합도를 정의하고자 할 때나, 적합도 자체가 컴퓨터 상에서 직접 정의하기 어려울 때 주로 사용된다. 물론 사용자로부터 평가를 받는 작업은 평가횟수가 많아질수록 사용자의 피로로 이어진다는 어려움이 있다. 따라서 사람에 의한 평가 방식은 대화형 진화연산의 주요한 강점인 동시에 성능을 저하시키는 요인이 된다. 실제로 이 문제를 해결하기 위한 몇 가지 방법이 제안되었고, 널리 사용되고 있다. 인공지능망을 이용하여 일부 평가를 자동화하는 방법이 여기에 속한다[1].

2.4 진화에 의한 창조

진화 알고리즘이 만들어내는 개체의 다양성을 이용하여 창의적인 미술과 음악 분야에 사용한 예가 있다. 복잡한 미술품을 만들어내려는 시도인 Todd 와 Latham 의

FormSynth[7]가 있고, 건축 표면을 생성하는 Hemberg 의 GenR8[9], 즉흥 재즈를 작곡하는 Biles 의 GenJam[3] 등이 있다.

본 논문에서는 진화를 이용하여 새로운 것을 창조하려는 시도의 한 영역을 개척하려는 것으로서, ‘창조적인 개념의 생성’을 다룬다고 할 수 있다. 진화를 통해 토픽맵을 새롭게 생성함으로써 개념을 재정의하거나, 개념들간의 새로운 관계를 만들어냄으로써 해당 지식에 대한 이해를 높이고 새로운 아이디어를 창출해낼 수 있을 것으로 기대된다.

3. 실험 설계

3.1 실험의 정의

본 실험은 토픽맵에 진화연산을 적용함으로써 새로운 창의성 도구를 만들고자 하는 시도의 첫 과정으로서, 주안점은 최종적으로 완성적인 형태를 제시하고자 하는 것이 아니고, 토픽맵에 진화연산을 적용할 수 있는지에 여부를 최초로 실험하는 데에 있음을 밝혀 둔다. 그렇기 때문에 현재로서는 토픽맵의 표준을 정확히 지키거나, 사용성이 높은 시각화와 그래픽 사용자 인터페이스를 제공하고 있지는 않다는 점에 유의하도록 한다.

따라서 본 논문에서 주요한 실험의 내용은 토픽맵 진화에 필요한 유전자형을 정의하고, 이를 바탕으로 구체적인 진화연산을 정의하여 토픽맵을 진화시킬 수 있는 간소한 인터페이스를 제공하는 것이라고 정의한다. 이를 바탕으로 앞으로의 연구에서 더욱 발전된 형태의 토픽맵 진화 시스템을 만들 수 있을 것이다.

3.2 실험 전개

3.2.1 실험 및 개발 환경

본 실험에서는 토픽맵을 진화하기 위한 시스템으로서 유전 알고리즘을 사용하였으며, 구체적으로 확장이 용이한 C++ 유전알고리즘 라이브러리인 GAlib(<http://lancet.mit.edu/ga>)을 기반으로 리눅스 환경(kernel 2.6)에서 구현되었다.

3.2.2 시스템 설계

본 토픽맵 진화 시스템의 설계는 크게 돌연변이, 교배, 선택 등과 같이 진화 연산자가 적용되는 부분과 사람이 개입되는 평가(evaluation)부분으로 나뉜다. 진화 시스템은 다른 유전 알고리즘의 순환 모델과 같은 형태를 가지며, 이는 다음 그림 1에 나타나 있다.

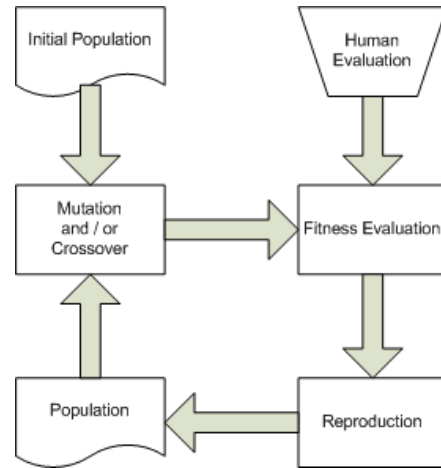


그림 1 시스템 설계

현재 시스템에서는 사람이 직접 모든 적합도 평가에 개입해야 하므로, 수십 내지 수백 개의 개체를 역시 수백 이상의 세대에 걸쳐 진화를 시킨다는 것은 어려움이 있다. 따라서 본 실험에서는 10 개의 개체를 10 여 번의 세대에 걸쳐 평가하는 것으로 제한하였다. 이와 같은 실험 상수는 다음 표 1에 정리하였다.

이름	값
인구수	10
세대	10
단어수(DICSIZE)	200~2000
돌연변이 확률(mutrate)	$0.1 \div \text{DICSIZE}$
노드 돌연변이 확률	$\text{mutrate} \times 0.2$
간선 돌연변이 확률	$\text{Mutrate} \times 0.8$

표 1 실험 상수

3.2.3 유전자형(genotype) 정의

토픽맵 표준에 의하면 토픽맵은 주제(topic), 관계(association), 그리고 발현(occurrence) 등의 요소로 이루어져 있지만, 이 표준을 다소 느슨하게 적용시키면 일종의 마인드맵 혹은 그래프로 추상화될 수 있는데, 구체적으로는 유향성 그래프(directed graph)로 축소하여 표현할 수 있다. 즉, 이렇게 표현된 토픽맵은 단순한 마인드맵(mind map)처럼 노드(개념), 노드와 노드를 연결하는 (이름을 가진) 간선으로 구성된다.

본래 유향성 다중그래프(directed multigraph)에 가까운 토픽맵을 유향성 그래프로 축소하여 얻을 수 있는 효과는 비교적 사람이 이 시스템을 통해 새로운 개념을 발견하고 만들어낸다는 목적에 크게 어긋나지 않으면서도,

진화연산을 시작하는 데에 있어 복잡한 유전자형을 구현하는 어려움을 크게 줄여줄 수 있다는 장점이 있다.

유향성 그래프를 구현하는 방법에는 여러 가지가 있고, 대체로 유전 알고리즘과 같은 진화연산기법에서는 여러 번의 세대를 실행하고, 수십 내지 수백 개의 개체를 평가해야 하므로 속도가 빠르고, 공간 복잡도도 그리 크지 않은 자료구조 및 알고리즘이 요구되지만, 본 실험에서는 컴퓨터의 연산 부분 보다 인간의 평가 부분이 더욱 많은 병목을 차지하는 시스템을 구현하게 되므로 구현상의 편의가 더욱 많이 고려될 수 있었다.

본 실험에서는 이러한 유향성 그래프를 인접 매트릭스(edge matrix)를 이용하여 구현하였다. 총 수백~수천여 개에 달하는 특정 분야의 단어들을 대상으로 하고, 이들 단어(혹은 개념)가 토픽맵상에 존재하는지 여부는 인접 매트릭스 상에 간선을 하나라도 가지는지 여부에 의해 결정된다.

3.2.4 진화 연산자(genetic operator) 정의

앞서 설계한대로 표현 방법을 이용하여 진화연산에 필요한 유전자형을 정의하고, 이를 바탕으로 돌연변이(mutation)와 교배(crossover)와 같은 진화 연산자(evolutionary operators)를 정의한다. 단, 본 실험에서는 교배 연산의 복잡성 때문에 이 부분은 본문에서 논하되 구현은 앞으로의 과제로 남겨둔다.

돌연변이 연산은 노드와 간선에 대한 임의의 변형으로 정의되며, 이는 인접 매트릭스 상에서 손쉽게 구현이 가능하다. 노드의 추가, 삭제, 변형 등이 있을 수 있고, 간선의 추가, 삭제 등이 가능하다.

사람의 평가 부분은 그래픽 사용자 인터페이스가 아닌 커맨드라인 인터페이스(CUI)로 구성되었다. 사람이 직접 토픽맵을 수치로 평가할 수 있으며 여기서 토픽맵은 그래프를 단순 텍스트로 표현하여 나타낸다.

3.2.5 기타 사항

마지막으로 본 실험에서는 예술에 관한 2000 여 개의 영어 단어를 사전으로 하였고, 이들 중 일부 혹은 전부를 단어 집합으로 하여 진행하였다.

4. 본론

4.1 실험 결과 분석

위 3 절에서 정의한 시스템을 구현하여 실험해본 결과, 다음과 같은 점이 발견되었다.

우선, 사용되는 사전의 크기가 너무 클 경우, 진화를 구체적으로 조절하기가 쉽지 않다는 점이 확인된다. 즉,

너무나 다양한 단어 집합으로부터 토픽맵을 만들 경우, 좋은 교배 방법이 구현되기 전까지는 돌연변이만으로 진화를 조절하기가 쉽지 않다. 사용자가 직접 자유 연상한 수십 개의 단어만으로 토픽맵을 그려보는 것이 사람이 보기에 자연스러운 형태로 수렴할 수 있을 것으로 보인다.

두 번째는 진화연산상의 상수 설정(parameter setting)에 관한 것으로서, 노드와 간선에 대한 돌연변이 확률이 역시 사람이 보기에 부담스럽지 않고 자연스럽게 이루어질 수 있는 수준으로 조절되어야 한다는 점이 발견된다. 돌연변이가 한꺼번에 너무 많이 이루어지거나, 자연스럽게 않게 이루어질 경우, 사람의 평가 자체가 진화상에 의미를 부여하기 어려워짐이 발견된다.

마지막으로 본 시스템이 새로운 아이디어를 생성할 수 있는 가능성을 보여주는 하지만, 사람이 각 개체를 평가하기 때문에 매우 느리고 피로로 인해 진행이 지속적이지 않다는 점을 확인할 수 있었다.

4.1 논의

일단 이 시스템은 돌연변이만으로 구성되었지만, 새로운 토픽맵을 만들어내면서 사람이 부여한 적합도를 따라 점점 더 발전해가는 토픽맵을 진화시킬 수 있음을 충분히 보여주고 있다.

돌연변이 연산은 상대적으로 간단하게 정의할 수 있다. 토픽맵상의 돌연변이는 한 개 혹은 인접한 여러 개의 노드 및 간선의 내용을 무작위로 확장, 삭제, 변경과 같은 작업을 가하는 것이다. 반면에 교배 연산을 위해서는 이보다 복잡한 과정으로 요하며, 유사도 메트릭(distance metric)이 잘 정의되는 것이 중요하다. 그래프에 대해 먼저 하위그래프의 삽입, 삭제, 변경, 이동과 같은 변환을 만들고 두 그래프가 같아지기 위해서 이와 같은 변환이 얼마나 많이 필요한지 측정할 수 있다. 이를 바탕으로 유사도 메트릭을 정의할 필요가 있다. 이렇게 정의된 유사도 메트릭을 이용하여 교배 연산을 만들 수 있고, 돌연변이 연산과 함께, 기대되는 진화 연산의 효과를 이끌어 낼 수 있을 것이다. 다만 본 논문에서는 양질의 교배 연산을 정의하는 어려움이 있으므로 앞으로의 이루어질 과제로서 남겨둔다.

비록 본 논문에서 보인 토픽맵 진화 시스템이 새로운 토픽맵을 진화시킬 수 있음을 보여주고는 있지만, 사람이 모든 개체를 평가하는 데에서 오는 병목과 피로의 문제는 앞으로 반드시 해결되어야 할 것으로 보인다. 좀 더 많은 개체를 두고 더욱 많은 세대를 발전시키면서 완성적인 형태의 토픽맵을 생성해낼 수 있을 것으로 기대되기 때문이다.

5. 결론 및 앞으로의 과제

대화형 진화연산 방법을 토픽맵에 적용하여 사람이 쉽게 만들어낼 수 없는 새롭고 창의적인 새로운 지식을

만들어낼 수 있다. 본 연구에는 사람의 평가가 매 단계마다 이루어지는 방식을 사용하였다. 이 시스템으로부터 새로운 토픽맵을 생성되고 이로부터 새로운 지식에 대한 통찰을 얻을 수 있을 것이다. 나아가서, 이렇게 만들어진 토픽맵이 창의성을 도모하는 도구로서 사용될 수 있을 것이다.

본 연구의 토픽맵 진화와 같이 창의성을 이끌어내기 위한 시스템에서는 단일한 적합도 높은 해를 구하는 것보다, 질이 좋으면서도 다양한 여러 해를 구하는 것이 더욱 중요하다. 따라서 이러한 점에 입각하여, 기존의 진화 연산 알고리즘을 수정, 단일한 적합도 기준에 영향을 받는 것이 아닌, 다양성에 더욱 비중을 둘 수 있는 형태로 발전시켜 보고자 한다.

또한 본 연구를 바탕으로 앞으로는 보다 사용성이 높은 시각화와 그래픽 사용자 인터페이스(GUI)를 구성해본다. 새로운 그래픽 사용자 인터페이스는 사용자로 하여금 보다 토픽맵을 평가하기 용이하게 하고, 그 평가 방식도 단순 평가 수치가 아닌 토픽맵의 특정 부분에 대한 구체적인 평가가 가능한 형태로 제시될 수 있을 것이다.

한편 인공신경망(ANN)과 같은 학습 알고리즘을 통해 토픽맵의 진화 과정 중에 개입되는 사람의 정신적, 육체적 수고를 덜 수 있는 방법을 고안할 것이다. 또한 토픽맵 평가에 있어서 단순 수치만을 사용하는 것보다 높은 수준의 평가 언어를 도입할 수 있을 것이다.

마지막으로, 시간과 연구 자원상의 여력이 있다면 사용자 인터페이스와 사용성을 향상시키는 이러한 노력을 보다 객관적으로 평가하기 위해 다수의 사용자에 대해 심층적인 분석을 수행해볼 것이다.

References

[1] Johanson, B. and Poli, R. 1998. GP-music: An interactive genetic programming system for music generation with automated fitness raters. In John R. Koza, Wolfgang Banzhaf, Kumar Chellapilla, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max H. Garzon, David E. Goldberg, Hitoshi Iba, and Rick Riolo,

editors, Proceedings of the Third Annual Conference on Genetic Programming, pp. 181-186.

[2] Le Grand, B. and Soto, M. 2000. Information management: Topic Maps visualization, XML Europe 2000, Paris, France.

[3] Biles, J. A. (1994). GenJam: A Genetic Algorithm for Generating Jazz Solos. In Proceedings of the 1994 International Computer Music Conference, IGMA, San Francisco.

[4] Pepper, S. 2000. The TAO of Topic Maps. In Proceedings of XML Europe 2000, Paris, France. <http://www.ontopia.net/topicmaps/materials/rdf.html>

[5] Takagi, H. 2001. Interactive Evolutionary Computation: Fusion of the Capacities of EC Optimization and Human Evaluation. Proceedings of the IEEE 89, 9, pp. 1275-1296.

[6] C. R. Darwin 'On the Origin of Species by Means of Natural Selection', Murray, London 1859

[7] S. Todd and W. Latham 'Evolutionary Art and Computers', Academic Press, 1992, ISBN 012437185X

[8] International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), Topic Maps, International Standard ISO/IEC 13250:1999, April 19, 1999.

[9] M. Hemberg, U-M. O'Reilly and P. Nordin, 'GENR8 - A Design Tool for Surface Generation', GECCO 2001 late-breaking-papers proceedings

[10] J. H. Holland, 'Adaptation in Natural and Artificial Systems', U. Michigan, 1975

[11] D. E. Goldberg, 'The Design of Innovation: Lessons from and for Competent Genetic Algorithms', Addison-Wesley, Reading, 2002