한국지능시스템학회 논문지 2010, Vol. 20, No. 2, pp. 234-242

정점 간의 거리 보존 및 최소 간선 교차에 기반을 둔 유전 알고리즘을 이용한 그래프 시각화

Graph Visualization Using Genetic Algorithms of Preserving Distances between Vertices and Minimizing Edge Intersections

계주성 · 김용혁* · 김우생

Ju-Sung Kye, Yong-Hyuk Kim* and Woo-Sang Kim

광운대학교 컴퓨터소프트웨어학과

요 약

본 논문은 가장 중요한 자료구조 중 하나인 그래프를 효과적으로 시각화하는 방법에 관해 다룬다. 그래프 시각화의 목적은 원래의 그래프 구조를 이해하기 쉽게 잘 표현하는 것이지만 일반적으로 그래프의 크기가 커짐에 따라 교차하는 간선의 수가 급격히 증가하여 각 정점 간의 관계를 눈으로 확인하기 쉽지 않게 된다. 우리는 이 문제에 대한 새로운 해결 방법을 제시한다. 기존의 연구들이 대부분 간선의 교차 수를 최소화하는 조건만 고려하였던 것을 벗어나 본 논문에서는 간선 교차 최소화와 더불어 정점 간의 거리를 보존하는 것을 동시에 고려하여 원래의 그래프 구조를 최대한 잘 표현하고자 하였다. 이를 위하여 본 논문에서는 이 두 목적 값의 가중치 합으로 각 해를 평가하는 유전 알고리즘을 설계하여 시각화에 적용한다. 제안한 시각화 방법이 새로운 목적에 맞게 그래프를 잘 시각화할 수 있음을 실험을 통해 입증할 수 있었다.

키워드 : 그래프 시각화, 유전 알고리즘, 새몬(Sammon)의 매핑

Abstract

In this paper, we deal with the visualization of graphs, which are one of the most important data structures. As the size of a graph increases, it becomes more difficult to check the graph visually because of the increase of edge intersections. We propose a new method of overcoming such problem. Most of previous studies considered only the minimization of edge intersections, but we additionally pursue to preserve distances between vertices. We present a novel genetic algorithm using an evaluation function based on a weighted sum of two objectives. Our experiments could show effective visualization results.

Key Words : Graph drawing, graph visualization, genetic algorithms, Sammon's mapping.

1. 서 론

그래프 시각화(graph visualization)는 여러 데이터 요소 간의 관계를 알기 쉽게 표현해주는 방법으로서, 컴퓨터 시스템의 파일 계층구조, 생물학 분야의 생태의 진화분류와 계통분류, 화학 분야의 분자구조 표현 등의 다양한 분야에서 응용되어 사용되고 있다. 시각화가 잘된 그래프란 사용자가 그 구조를 이해하기 쉽게 그려진 그래프를 말하는데, 그동안 이 분야에 대한 많은 연구들이 이루어져 왔다. 하지만 이해하기 쉬운 그래프를 그리기 위해 고려해야 할 요소들이 많기 때문에, 몇 개의 정점(vertex)과 간선(edge)들로 이루어진 그래프라면 그리 어렵지 않게 시각화를 잘 할 수 있지만, 이보다 훨씬 더 많은 정점과 간선들로 구성된 그래프를 보기 좋게 잘 시각화하기란 만만치 않다. 그래프 시각화를 위해 고려되는 주된 요소는 크게 구조적(structural) 측면과 의미적(semantic) 측면의 두 가지로 나누어진다[1]. 먼저 구조적 측면의 요소는 간선이 겹쳐지는 것을 피하고 일정한 규칙성을 가지도록 배치하는 것에 초점이 맞춰져 있고, 의미적 측면의 요소는 그래프 시각화 방법으로 표현한 자료들이 원래 가지고 있는 중요한 특징을 보존하는 것에 초점이 맞춰져 있다.

본 논문은 그래프 시각화의 구조적 측면의 요소 중에서 가장 중요한 요소로 여겨지는 간선 교차 최소화의 문제, 즉, 이차원 평면상에서 그래프를 시각화할 때 정점과 간선 수가 많아지면 교차하는 간선의 수가 증가하여 각 데이터 간의 관계를 눈으로 확인하기 어렵다는 문제에 대한 해결 방법을 제안한다. 기존의 방법들은 대부분 간선의 교차 최소화라는 하나의 조건만을 고려하였지만, 본 논문에서는 정점 간의 거리 보존과 간선 교차 최소화, 두 가지 조건을 동시에 고려하여 원래의 구조를 크게 손상하지 않으며 시각화하고자 한다. 문제 해결 방법으로 정점 간의 거리를 보존하고 간선 교차를 최소화하는, 두 목적 값의 가중치 합으로 수식화된 평가함수에 기반을 둔 유전 알고리즘[2](GA: genetic algorithm)을 설계하고, 실험을 통해 이 방법이 두 가지 조건을

접수일자 : 2009년 3월 23일
완료일자 : 2009년 9월 16일
* 교신 저자
본 논문은 2009년도 광운대학교 연구년 지원에 의해서 연구되었음.

균형 있게 반영하여 그래프를 이해하기 쉽게 잘 시각화할 수 있음을 보이고자 한다.

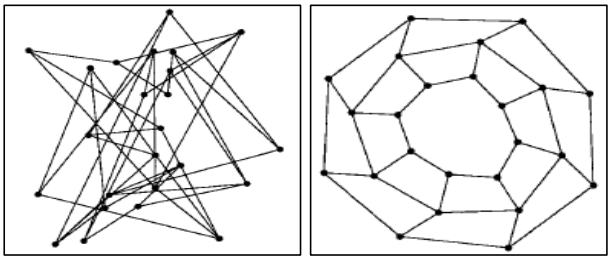
2. 관련 연구

2.1 그래프 시각화

그래프 시각화의 목적은 데이터 간의 관계를 눈으로 이해하기 쉽게 표현하는데 있다. 시각화가 잘된 보기 좋은 그래프를 위해 고려되는 요소로는 정점의 위치, 간선의 위치, 그래프의 전체적인 레이아웃 등으로 크게 나눌 수 있다. 그렇다면 보기 좋은 그래프란 어떤 그래프일까? Batttista는 여러 관련 논문을 정리하면서 어떤 그래프가 보기 좋은 그래프인지에 대한 몇 가지 규칙(rule)을 제시하였다[3]. 그 규칙을 정리하면 아래와 같다.

- 정점과 간선은 **균등하게 분산**되어야 한다.
- 간선은 모두 **같은 길이**를 가져야 한다.
- 간선은 **직선**이어야 한다.
- **동일한 구조**의 그래프는 같은 방법으로 그려져야 한다.
- 간선의 **교차가 최소화**되어야 한다.

이 조건들은 서로 상반관계(trade-off)에 있기 때문에 어느 한 가지 요소의 특징을 좋게 하면 다른 요소들의 특징이 저하된다. 그렇기 때문에 원래 데이터 간의 특징을 잘 표현할 수 있게 위 요소들의 적절한 조화 점을 찾는 것이 중요하다[4]. 그림 1은 같은 그래프에 대해 위 조건들과 관련해 시각적으로 잘 그려지지 않은 예와 잘 그려진 예를 비교한 것이다. 왼쪽 그래프의 경우는 다수의 간선이 교차하고 있고, 정점과 간선이 균등하게 분산되어 있지 않았다. 또한 간선의 길이가 일정치 않아 보기에 좋지 않다. 반면에 오른쪽 그래프는 간선의 교차가 완전히 제거되었고 정점과 유사 길이의 간선이 규칙적으로 배치되어 있어 훨씬 더 보기 좋은 것을 알 수 있다.



(a) 시각화가 잘 안된 예 (b) 시각화가 잘된 예

그림 1. 그래프 시각화의 예
Fig. 1. An example of graph visualization

그렇다면 시각적으로 보기 좋은 그래프를 위한 가장 중요한 조건은 무엇일까? Purchase는 여러 연구[5-7]를 통해 시각화가 잘된 좋은 그래프의 조건들 가운데 우선순위를 매겼는데, 그 중 가장 중요한 요소로 간선의 교차가 최소화되어야 한다고 결론을 내렸다. 이 간선 교차의 최소화 문제는 복잡하고 어려운 문제인데 Garey와 Johnson[8]은 이 문제가 NP-hard임을 증명하였고, Eades와 Whitesides[9]는 NP-complete라는 것을 증명하였다. 그리고 이 문제를 해결을 위한 Tutte[10]의 Baycenter 휴리스틱 방법을 시작으로 다양한 간선 교차의 최소화를 위한 휴리스틱 방법들이 연구되어 왔다[11-13].

2.2 유전 알고리즘

유전 알고리즘은 다윈(Darwin)의 진화원리를 모방하여 종의 선택과 교차를 통하여 염색체의 적합도(fitness)를 평가하고, 선택된 우수 개체를 바탕으로 새로운 세대를 만들어 내며 탐색해가는 메타휴리스틱 최적화 기법이다. 유전 알고리즘은 크게, 문제의 해를 표현하는 염색체(chromosome), 염색체로 다수로 구성된 해 집단(population), 해의 적합도(fitness)를 평가하는 평가 함수(evaluation function)로 구성되며, 이 염색체들을 선택(selection), 교차(crossover), 변이(mutation)의 과정을 거쳐 점진적으로 개선시켜 나감으로써 점차 최적에 가까운 해를 얻게 된다[14]. 유전 알고리즘의 전체적인 흐름은 그림 2와 같고, 각 단계에 대한 설명은 다음과 같다.

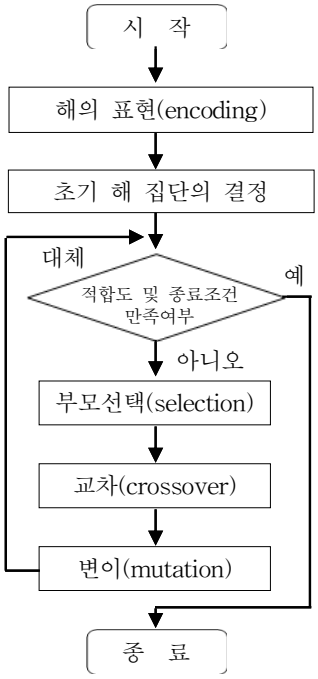


그림 2. 유전 알고리즘의 일반적인 구조
Fig. 2. The general structure of genetic algorithm

어떤 문제를 푸는데 유전 알고리즘을 적용하기 위해서는 문제의 해를 적절히 인코딩하여 표현해야 한다. 전형적인 표현방법으로는 '0'과 '1'을 이용한 선형적(linear) 이진 표현(binary encoding)이 있고, 이 밖에도 그레이 코딩(gray coding), 실수 표현, 가변길이 표현, 등 여러 가지 방법이 존재한다. 이렇게 문제의 해를 염색체로 표현하여 일정 개수의 각각 다른 염색체로 초기 해 집단을 구성하게 되고, 이 해 집단에 여러 유전 연산자를 적용하여 해를 개선시켜 나가게 된다. 유전 연산자에는 선택, 교차, 변이 연산이 있다. 먼저 선택 연산은 초기 유전자 집단에서 교차를 위해 두 개의 부모 염색체를 선택하는 과정으로 일반적으로 우수한 품질의 해가 선택될 가능성을 높게 한다. 두 개의 부모 염색체가 선택되면 두 부모 염색체를 부분적으로 추출하여 자식

해를 얻는 교차 연산을 적용하게 된다. 이렇게 생성된 자식 해는 해의 다양성을 위해 부모 해에 없는 속성을 생성하는 변이 연산을 거쳐 최종적인 자식 해가 만들어진다. 얻어진 자식 해는 보통 해 집단에서 가장 좋지 않은 해와 교체된다. 유전 알고리즘은 지금까지의 과정을 종료조건을 만족될 때까지 반복하게 된다. 종료 조건은 다양한 방식이 있지만 대개의 경우 적합도를 판단하여 해의 품질이 일정 정도에 이르게 될 때 종료하거나 특정 세대 수만큼 진화 연산을 적용한 후 종료하는 방식을 취한다.

2.3 지역 최적화

새몬 매핑(Sammon's mapping)[15]은 고차원 데이터를 저차원 공간으로 매핑하는 방법으로 기본 아이디어는 고차원의 입력 공간의 모든 점들의 각 관계 정도의 변형을 최소화하여 저차원의 공간으로 매핑하는 것이다. 새몬 매핑은 데이터 간의 거리를 최대한 보존하려고 하는데, 처음의 입력 공간과 출력 공간에서의 데이터의 거리의 차를 비교하여 에러율(stress measure)을 최소화하고자 한다. 데이터 집합의 개수를 n이라 하고, 입력 공간의 두 점 x_i와 x_j 사이의 거리를 δ_{ij}라고 하고, 매핑되는 출력 공간의 두 점 x_i', x_j' 사이의 거리를 d_{ij}라고 했을 때, 새몬 매핑의 에러율 M은 아래 식 (1)과 같다[16].

$$M = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{(\delta_{ij}-d_{ij})^2}{\delta_{ij}} \quad (1)$$

M 값의 범위는 0에서 1 사이이며 결과 값이 작을수록 정보 손실이 적어 매핑이 잘 되었음을 의미한다.

3. 그래프 시각화를 위한 유전 알고리즘

정점의 거리를 보존하고 간선의 교차 수를 최소화하기 위해 본 논문은 유전 알고리즘을 채택한다.

```
n 개의 초기 염색체 생성;
생성된 염색체들로 해 집단 구성;
해 집단의 각 해를 평가함수로 평가;
repeat {
    해 집단에서 두 부모 해 p1, p2 선택;
    부모 해 p1, p2를 교차하여 자식 해 생성;
    자식 해를 변이시킴;
    변이된 자식 해에 지역 최적화 적용;
    자식 해를 평가함수로 평가한 후 해 집단 내의 한
    염색체와 대체함;
} until (종료조건);
return 해 집단에서 가장 좋은 해;
```

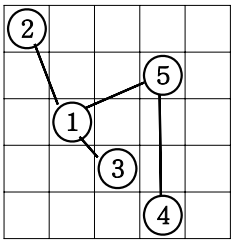
그림 3. 혼합형 안정상태 유전 알고리즘의 구조
Fig. 3. The structure of hybrid steady-state genetic algorithms

우리는 유전 알고리즘의 여러 방법론 중 지역 최적화 방법을 결합한 혼합형(hybrid) 안전상태(steady-state) 유전 알고리즘을 사용했다. 안정상태 유전 알고리즘은 한 세대동안 해 집단에서 하나의 해만 바뀌는 방식이고 혼합형 유전 알고리즘은 자식 해 생성에 지역 최적화 알고리즘을 결합하여 해 집단을 진화시켜 나가는 방식이다. 제안한 유전 알고리즘의 구조는 그림 3과 같이 구성된다. 유전 알고리즘의 각 단계에 대한 설명은 이 장의 소절에서 각각 설명하기로 한다.

3.1 해의 표현

유전 알고리즘을 사용하기 위해서는 먼저 그래프를 염색체로 표현해야 한다. 일반적인 그래프는 각 정점들 간의 인접 여부 정보를 포함하는 인접행렬(adjacent matrix)로 표현된다. 하지만 그래프의 시각화를 위해서는 그래프의 각 정점을 공간상에 표시해야 하기 때문에 정점의 좌표정보가 더 필요하다. 각 정점의 좌표는 이차원의 랜덤으로 생성하였고, 이 좌표를 정점 별로 묶어 염색체를 표현하였다.

그림 4는 그래프 정점의 위치 정보를 염색체로 표현한 예를 보여준다. 그래프의 정점들이 이차원 공간에 위치해 있고(좌측 상단 그림) 주어진 그래프의 정점들 간의 인접여부를 0과 1로 표현한 인접행렬이 있다면(우측 상단 그림), 인접여부와 상관없이 각 정점들의 위치 정보만을 사용하여 해를 표현할 수 있다. 예를 들면, 정점 1의 경우는 정점의 x, y의 좌표가 (2, 3)이므로 염색체의 첫 번째 유전자는 (2, 3)이 된다. 나머지 네 개의 정점도 이러한 방법으로 구성하면 하나의 염색체가 만들어진다.



<그래프의 이차원 배치>

0	1	1	0	1
1	0	0	0	0
1	0	0	0	0
0	0	0	0	1
1	0	0	1	0

<인접행렬>

(2, 3)	(1, 5)	(3, 2)	(4, 1)	(4, 4)

<염색체>

그림 4. 그래프와 염색체 표현의 예
Fig. 4. An example of a graph and its representation

이와 같은 방법으로 랜덤 좌표를 가지는 n 개의 정점으로 구성된 염색체를 m 개를 만들어 그림 5와 같이 초기 해 집단을 구성한다.

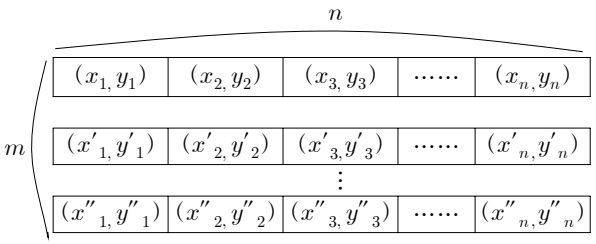


그림 5. 해 집단
Fig. 5. Population

3.2 평가 함수

각 해의 적합도를 구하기 위해 그 품질을 평가해야 하는데, 품질의 평가 항목에는 정점 간의 거리의 보존 정도와 간선의 교차점 개수가 포함된다. 아래의 함수 식 (2)는 위의 두 가지 요소를 고려한 평가함수이다. 함수에서 C는 간선의 교차점의 개수이고, M은 2.3 절의 새몬 매핑 에러율 값으로 정점 간의 거리를 얼마나 보존하였는가에 대한 평가 값이다. 새몬 매핑에서 사용된 간선의 거리는 Floyd 알고리즘을 사용해 계산된 각 정점 간의 최단 경로 길이(shortest path length)를 이용하였다. 여기에 α, β 상수 값으로 각각에 가중치를 부여하여 간선의 교차점 개수와 정점 간의 거리 보존 값, 이 두 가지 요소가 평가함수에서 비슷한 비중을 차지할 수 있게 조정하였다.

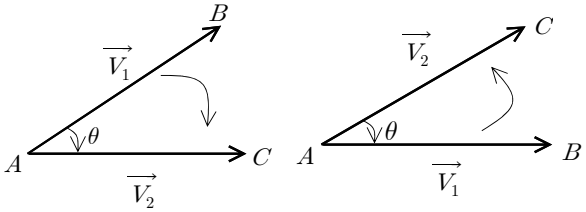
$$F = \alpha C + \beta M \tag{2}$$

· 간선의 교차여부 판별 : 평가함수에 포함되어 있는 간선의 교차점의 개수를 구하기 위한 간선의 교차를 판별하는 방법은 다음과 같다. 먼저 단순한 방법으로는 간선은 두 점, (x_1, y_1)과 (x_2, y_2)로 정의될 수 있는데, 서로 다른 두 점을 지나는 두 간선의 기울기가 서로 평행이 아니라면 두 간선은 반드시 교점이 있다. 이 때 교점의 좌표가 각 간선 위에 있는지를 판단하여 두 간선이 교차하는지 여부를 판단할 수 있지만 이 방법은 과정이 복잡하고 계산량이 많기 때문에 많은 수의 간선의 교차를 판단하는데 사용하기에는 효율적이지 않다. 그래서 좀 더 계산을 줄인 벡터의 외적(cross product)을 이용한 방법을 사용하였다.

$\overrightarrow{V_1}$, $\overrightarrow{V_2}$의 외적은 두 벡터에 수직인 새로운 벡터로서 크기는 $|\overrightarrow{V_1}\|\overrightarrow{V_2}|\sin\theta$가 되며, 방향은 $\overrightarrow{V_1}$과 $\overrightarrow{V_2}$에 수직인 두 벡터 중에서 한쪽 방향이 선택된다. 교차를 판단하기 전에 사전 연구로 그림 6은 xy-평면상에 세 점 A, B, C가 있을 때, A, B, C 세 점을 차례로 이을 때의 방향을 나타낸 것이다. $\overrightarrow{V_1}$을 A에서 B로 가는 벡터로 잡고, $\overrightarrow{V_2}$를 A에서 C로 가는 벡터로 잡으면 $\overrightarrow{V_1}\times\overrightarrow{V_2}$는 $\overrightarrow{V_1}$과 $\overrightarrow{V_2}$에 수직인 두 벡터 중 한쪽 방향이 선택되는데, (a)와 같이 시계 방향일 때 오른손 법칙에 의해 엄지손가락이 가리키는 방향을 z 축의 양의 방향(+)이라고 생각한다면 (b)와 같이 반시계 방향일 때 오른손 법칙에 의해 엄지손가락이 가리키는 방향을 z 축의 음의 방향(−)이라 생각할 수 있다. 이 때 $\overrightarrow{V_1}$, $\overrightarrow{V_2}$의 외적을 구해보면 식 (3)과 같다.

$$\begin{aligned}
&\overrightarrow{V_1} = \overrightarrow{B} - \overrightarrow{A} = (B_x - A_x,\ B_y - A_y,\ 0)\\
&\overrightarrow{V_2} = \overrightarrow{C} - \overrightarrow{A} = (C_x - A_x,\ C_y - A_y,\ 0)\\
&\overrightarrow{V_1}\times\overrightarrow{V_2} = (0,\ 0,\ (B_x - A_x)(C_y - A_y) - (B_y - A_y)(C_x - A_x))\\
&= (0, 0, B_x C_y - A_y B_x - A_x C_y - B_y C_x + A_x B_y + A_y C_x)
\end{aligned} \tag{3}$$

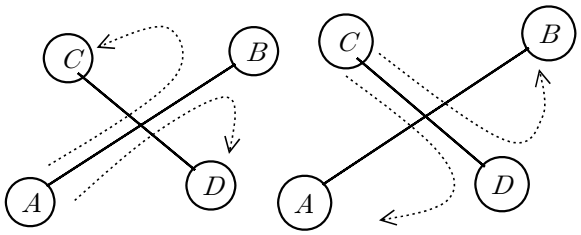
이와 같이 $\overrightarrow{V_1}$, $\overrightarrow{V_2}$의 외적은 x와 y 성분이 모두 0이고, z 성분만 있다. 즉, $\overrightarrow{V_1}$, $\overrightarrow{V_2}$의 외적의 z 성분이 음이면 $\overrightarrow{V_1}$, $\overrightarrow{V_2}$가 반시계 방향, 양이면 시계 방향이라고 판단할 수 있다.



(a) 시계 방향 (b) 반시계 방향

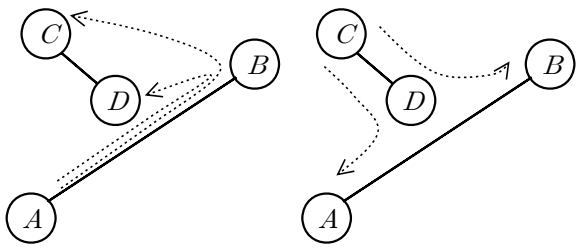
그림 6. 외적으로 방향 판단
Fig. 6. Determining direction by cross product

위의 방법을 이용하여 A, B, C를 차례로 선분으로 연결하여 그 방향이 시계 방향인지 아니면 반시계 방향인지 결과를 얻는 것을 간단하게 $P(A, B, C)$로 표시한다면, 결과값으로 만약 선분으로 연결할 때 방향이 시계방향이면 외적의 z 성분은 양수를 리턴하고, 만약 반시계 방향이면 외적의 z 성분은 음수를 리턴한다.

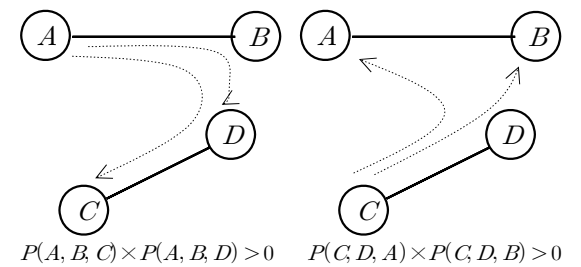


$P(A, B, C)\times P(A, B, D) < 0$ $P(C, D, A)\times P(C, D, B) < 0$

그림 7. 간선이 교차하는 경우
Fig. 7. Cases of edge-crossing



$P(A, B, C)\times P(A, B, D) > 0$ $P(C, D, A)\times P(C, D, B) < 0$



$P(A, B, C)\times P(A, B, D) > 0$ $P(C, D, A)\times P(C, D, B) > 0$

그림 8. 간선이 교차하지 않는 경우
Fig. 8. Cases of not edge-crossing

그림 7과 8은 그래프의 간선의 교차여부를 외적의 성질을 이용하여 판단한 것으로써 A, B 정점을 있는 선을 고정하고 이 선과 다른 두 정점 C, D와 각각 연결했을 때의 방

향을 판단하는 $P(A,B,C)\times P(A,B,D)$의 값과 반대로 C,D 정점을 잇는 선을 고정하고 정점 A,B와 각각 연결했을 때 방향을 판단하는 $P(C,D,A)\times P(C,D,B)$의 값을 고려하는데 만약 두 개의 간선이 교차한다면 그림 7과 같이 시계 방향과 반시계 방향을 갖게 되어 $P(A,B,C)\times P(A,B,D)$의 값과 $P(C,D,A)\times P(C,D,B)$ 값이 항상 음수가 된다. 만약 두 개의 간선이 교차하지 않는다면 그림 8과 같이 $P(A,B,C)\times P(A,B,D)$, $P(C,D,A)\times P(C,D,B)$값이 적어도 하나가 양수가 된다. 이 방법을 사용하면 적은 계산량으로 간선의 교차여부를 쉽게 판별할 수 있다.

3.3 선택 연산

평가함수로 평가된 해 중에서 교차에 사용할 두 개의 부모 해 p_1, p_2를 선택하는 방법으로는 가장 대표적인 룰렛휠(roulette-wheel) 선택 방법을 사용하였다. 룰렛휠 선택 방법은 다른 선택 방법과 마찬가지로 우수한 해가 선택될 확률이 높은 방법이다. 하지만 각각의 해의 품질을 그대로 선택될 확률을 결정하는 적합도로 사용할 경우 해 집단에서 가장 좋은 해와 가장 나쁜 해의 품질의 차이인 선택 압(selection pressure)이 높기 때문에, 나쁜 해들은 거의 선택될 기회를 잃게 된다. 그렇게 되면 해의 다양성이 급속히 떨어진다. 이러한 문제를 막기 위해 다음과 같은 적합도 계산 방법을 사용한다.

$$f_i=(C_w-C_i)+(C_w-C_b)/(k-1),\quad k=3$$

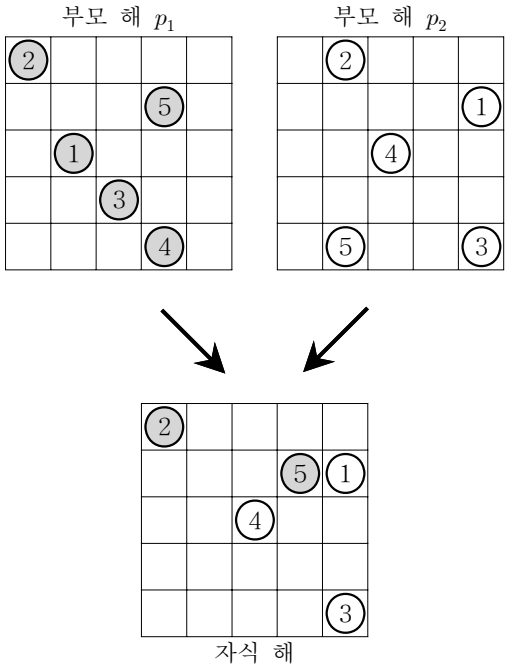
이 식에서 사용된 각 기호의 의미는 아래와 같다.

C_w : 해 집단 내에서 가장 나쁜 해의 평가함수 값
C_b : 해 집단 내에서 가장 좋은 해의 평가함수 값
C_i : 해 i의 평가함수 값

위의 방법을 사용하면 가장 좋은 해의 적합도가 가장 나쁜 해의 적합도보다 k 배가 되도록 조절할 수 있다. 본 논문의 실험에서는 일반적으로 가장 흔히 쓰는 k 값인 3을 사용하였다.

3.4 교차 연산

두 개의 부모 해 p_1, p_2가 선택이 되면 두 부모 해를 섞어 새로운 자식 해를 얻는 교차 연산을 수행하게 되는데, 교차 연산 방법은 균등 교차(uniform crossover) 방법을 사용하였다. 균등 교차는 먼저 임계 확률 p를 설정한 후 각각의 유전자 위치에서 난수를 발생한 다음 이 값이 p 이상이면 부모 해 p_1의 같은 위치로부터 유전자를 복사해오고, 그렇지 않으면 부모 해 p_2의 같은 위치로부터 복사를 한다. 아래 그림 9는 임계 확률 p가 0.5일 때 그래프에 균등교차 방법을 적용한 예를 보여준다.



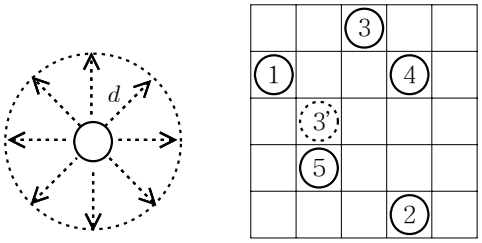
부모 해 p_1:	(2,3)	(1,5)	(3,2)	(4,1)	(4,4)
부모 해 p_2:	(5,4)	(2,5)	(5,1)	(3,3)	(2,1)
난수:	0.24	0.81	0.33	0.19	0.66
자식 해:	(5,4)	(1,5)	(5,1)	(3,3)	(4,4)

그림 9. 균등 교차 적용의 예 ($p=0.5$)
Fig. 9. An example of uniform crossover ($p=0.5$)

두 개의 부모 해 p_1, p_2가 선택되면 먼저 자식 해가 가지게 되는 첫 유전자를 p_1, p_2 중에서 선택하기 위해 0에서 1 사이의 난수를 발생시킨다. 여기서는 난수로 0.24가 생성이 되었고, 이것은 임계 확률 p보다 작기 때문에 자식의 첫 유전자는 부모 해 p_2의 처음 위치의 유전자를 가져오게 된다. 자식해의 두 번째 유전자의 경우에는 발생한 난수가 0.81로 임계 확률 p보다 크기 때문에 부모 해 p_1의 두 번째 위치에서 유전자를 가져오게 된다. 이 과정을 유전자의 수, 즉, 그래프 정점의 수만큼 반복하면 p_1, p_2가 교차된 자식 해를 얻을 수 있다.

3.5 변이 연산

변이 방법은 먼저 교차 연산과 마찬가지로 임계 확률 값 q를 설정하고, 0에서 1 사이의 난수를 발생해서 이 값보다 작은 경우의 유전자만 변이시킨다. 변이 내용은 유전자 정보인 그래프 정점의 위치를 정점의 중심에서 d 거리 내에 있는 새로운 위치로 이동시킨다. 그림 10은 임계 확률 q 값이 0.1이고 변화될 수 있는 거리 값 d가 2일 때 변이 적용의 예이다.



자식 해:	(1,4)	(4,1)	(3,5)	(4,4)	(2,2)
난수:	0.74	0.52	0.08	0.23	0.99
			⇓		
변이된 값:			(2,3)		

그림 10. 변이 적용의 예 (q= 0.1)
Fig. 10. An example of mutation (q= 0.1)

교차에 의해 생성된 자식 해에서 변이시킬 유전자를 선택하기 위해 각각의 유전자의 위치에서 난수를 발생시키게 되는데, 난수가 q 값인 0.1보다 작은 경우인 3 번째 정점만 변이가 된다. 변이 값은 변이 범위 값인 d = 2 내에서 (3,5)의 좌표가 (2,3)으로 변화되었다. 그리고 그 결과 간선의 교차점 개수가 줄었다. 정점 1과 4의 간선과 정점 2와 3의 간선이 겹치었는데, 정점 3의 위치가 3'로 바뀌면서 정점 1과 4의 간선과의 교차점이 사라졌다. 이와 같이 변이 연산은 보다 다양한 해의 생성에 기여해 해의 품질을 개선하는데 도움이 된다.

3.6 지역 최적화 및 대체 연산

지금까지의 연산들만 사용할 경우 만족할 만한 해를 얻을 때까지는 오랜 시간이 걸리게 되기 때문에 지역 최적화(local optimization) 방법을 함께 사용한다. 지역 최적화의 방법으로는 변이까지 한 세대가 끝난 해에 2.3 절의 새몬 매핑을 n 번 적용시켰다. 이렇게 함으로써 전체적인 해들의 품질이 좋아지게 되어 만족할만한 품질의 해를 빠른 시간 내에 구할 수 있다. 지역 최적화를 적용한 후 최종적인 자식 해는 지역 최적화를 하기 전의 해와 지역 최적화를 한 후의 해의 품질을 비교하여 이중 더 좋은 해가 최종 자식 해로 선정되게 된다.

최종 자식 해가 얻어지면 그 해를 다음 세대의 해 집단에 포함시킬지 여부를 판단하는 대체 연산을 하게 된다. 가장 손쉬운 방법은 해 집단 내에서 가장 품질이 낮은 해를 대치하는 것이다. 이는 유전 알고리즘의 수렴을 빠르게 하는 대신 조기 수렴(premature convergence)의 가능성이 크기 때문에 본 논문에서는 생성된 자식 해가 두 부모 해 중 어느 하나의 품질보다 좋은 경우에만 대체한다. 만약 자식 해가 두 부모 해 보다 더 품질이 좋은 경우에는 두 부모 해 중 품질이 더 낮은 부모 해와 대체하였다.

4. 실험 결과

실험에 사용한 그래프는 정점 간의 거리 보존과 간선 교차의 최소화의 결과를 쉽게 비교할 수 있는 그래프를 생성해서 테스트하였다. 총 세 가지 종류로 정점의 개수를 달리하여 S_{48}, S_{96}, S_{192}로 구분된다. 그래프 인스턴스 이름의 아래 첨자로 있는 숫자는 정점의 개수를 의미한다. 이 세 가지의 그래프를 랜덤, 순수 유전 알고리즘(pure GA), 새몬 매핑, 새몬 매핑을 지역최적화로 사용한 혼합형 유전 알고리즘(hybrid GA), 네 가지의 방법으로 나누어 실험을 하였고 실험에 사용한 유전 알고리즘의 파라미터 값 설정은 아래의 표 1과 같다.

표 1. 유전 알고리즘의 파라미터
Table 1. Genetic parameters

초기 해 : 랜덤(random) 생성
해 집단 : 30 개
변이 확률 : 0.1
변이 값 : 0.1
지역 최적화 : 새몬 매핑 30회
세대 수 : 50 세대
반속 횟수 : 50 회

해 집단은 해당 그래프의 정점을 임의의 초기 좌표로 설정한 그래프 30 개로 구성하였다. 그리고 변이 확률은 0.1로 변이가 일어났을 시에는 해당 정점의 좌표를 거리 0.1 내에서 이동하게 하였다. 지역 최적화는 GA와 새몬 매핑을 함께 사용한 혼합형 GA의 경우에만 적용하였고 자식 해에 새몬 매핑을 30 회 적용하여 지역 최적화하였다. 이와 같은 파라미터를 적용하여 각각의 그래프에 50 세대를 거쳐 유전 알고리즘이 종료된다. 이를 50 회 반복하여 평가 값이 가장 좋은 값(최상 값), 평균값, 표준편차를 구했다.

4.1 평가함수에서 가중치 결정

제대로 된 시각화를 수행하기 전에 3.2 절의 평가함수 식 (2)에 적용할 α, β 상수 값을 결정해야 한다. α, β 값에 따라 최종적으로 얻어지는 해가 간선의 교차점 수와 정점 간의 거리 보존 중, 어느 쪽으로 수렴하게 되는지에 대한 성향이 다르게 된다. 본 논문에서는 이 두 가지 요소를 다 고려하기 때문에 실험에 사용할 적절한 상수 값을 찾는 일은 필수적인 요소이다.

α, β 상수 값을 찾기 위해 실험에 사용한 그래프는 정점 수가 가장 적은 S_{48} 그래프를 사용하였고, GA 파라미터 값은 위의 표 1의 파라미터 값을 적용시켰다. 그 후 α, β 값 중 1 미만의 값을 갖는 새몬 매핑의 측정값에 적용되는 가중치인 β 값은 1로 고정시키고, 새몬 매핑의 측정값보다 상대적으로 큰 값을 갖는 간선의 교차 수에 적용되는 α 값은

표 2. 평가함수의 가중치 값 결정을 위한 실험 결과 (50 회 평균값)
Table 2. Experiments to decide weight factors in evaluation function (averages from 50 runs)

가중치 / 측정값	α =0.001	α =0.01	α =0.1	α =10	α =100
새몬 매핑 에러율(M)	0.0049	0.0055	0.0045	0.0048	0.0051
간선 교차점 수(C)	6.2	6.0	4.9	5.1	5.0

표 3. 실험 결과 (50 회 실행)
Table 3. Experimental results (from 50 runs)

그래프		랜덤			순수 GA			새몬 매핑			혼합형 GA		
		최상	평균	표준 편차	최상	평균	표준 편차	최상	평균	표준 편차	최상	평균	표준 편차
S_{48}	평가 값(F)	7.359	8.329	0.917	3.755	5.301	0.694	0.604	0.604	0.000	**0.405**	**0.495**	0.055
	교차점 수(C)	65	74.8	–	29	44.5	–	6	6.0	–	4	4.9	–
	SM 에러율(M)	0.859	0.851	–	0.855	0.851	–	0.004	0.004	–	0.005	0.004	–
S_{96}	평가 값(F)	13.974	15.828	1.252	10.780	12.731	0.981	1.804	1.804	0.000	**1.306**	**1.450**	0.081
	교차점 수(C)	131	150.1	–	99	118.0	–	18	18.0	–	13	14.5	–
	SM 에러율(M)	0.874	0.876	–	0.879	0.876	–	0.005	0.005	–	0.006	0.005	–
S_{192}	평가 값(F)	35.611	39.591	2.317	30.409	33.060	1.373	2.903	2.903	0.000	**2.404**	**2.639**	0.087
	교차점 수(C)	347	386.8	–	295	321.5	–	29	29.0	–	24	26.4	–
	SM 에러율(M)	0.911	0.911	–	0.909	0.910	–	0.003	0.003	–	0.004	0.004	–

β 값과의 비율을 0.001 배에서 100 배까지 바꿔가며 새몬 매핑 측정값과 간선의 교차수를 구하여 50 번 실행에 대한 평균값을 구하였다. 결과는 표 2와 같이 α 값이 0.1, β 값이 1일 때 가장 우수한 품질의 결과를 얻을 수 있었고, 추후의 모든 실험에 이와 동일한 상수 값을 적용시켜서 실험하였다.

4.2 실험 결과 및 결과 분석

표 3은 S_{48}, S_{96}, S_{192}의 세 개의 그래프에 대해 정점의 위치를 랜덤, 순수 GA, 새몬 매핑, GA와 새몬 매핑을 지역 최적화로 결합한 혼합형 GA, 네 가지 방법으로 시각화한 것의 결과 값을 보여준다. 세 그래프 모두 공통적으로 랜덤, 순수 GA, 새몬 매핑, 혼합형 GA 순으로 해의 품질이 좋다. 랜덤으로 정점을 배치한 경우에는 간선의 교차점 수나 정점 간의 거리를 조정하는 어떠한 기준도 없기 때문에 간선의 교차점의 수도 아주 많고, 정점 간의 거리도 전혀 보존되지 않아 품질이 매우 좋지 않은 것을 알 수 있다.

순수 GA를 사용한 경우에는 랜덤으로 정점을 배치한 것보다 교차점의 개수가 20~40% 정도 감소했다. 이것은 GA에 평가함수에 반영되어 있는 간선의 교차점 최소화 기준이 잘 적용되었기 때문이다. 반면에 새몬 매핑 측정값은 거의 비슷한 것을 볼 수 있는데, 그 이유는 비록 평가함수에서 새몬 매핑 측정값을 최소화하는 것을 고려하고 있지만 간선의 교차점의 수가 많기 때문에 평가 함수에서 차지하는 비

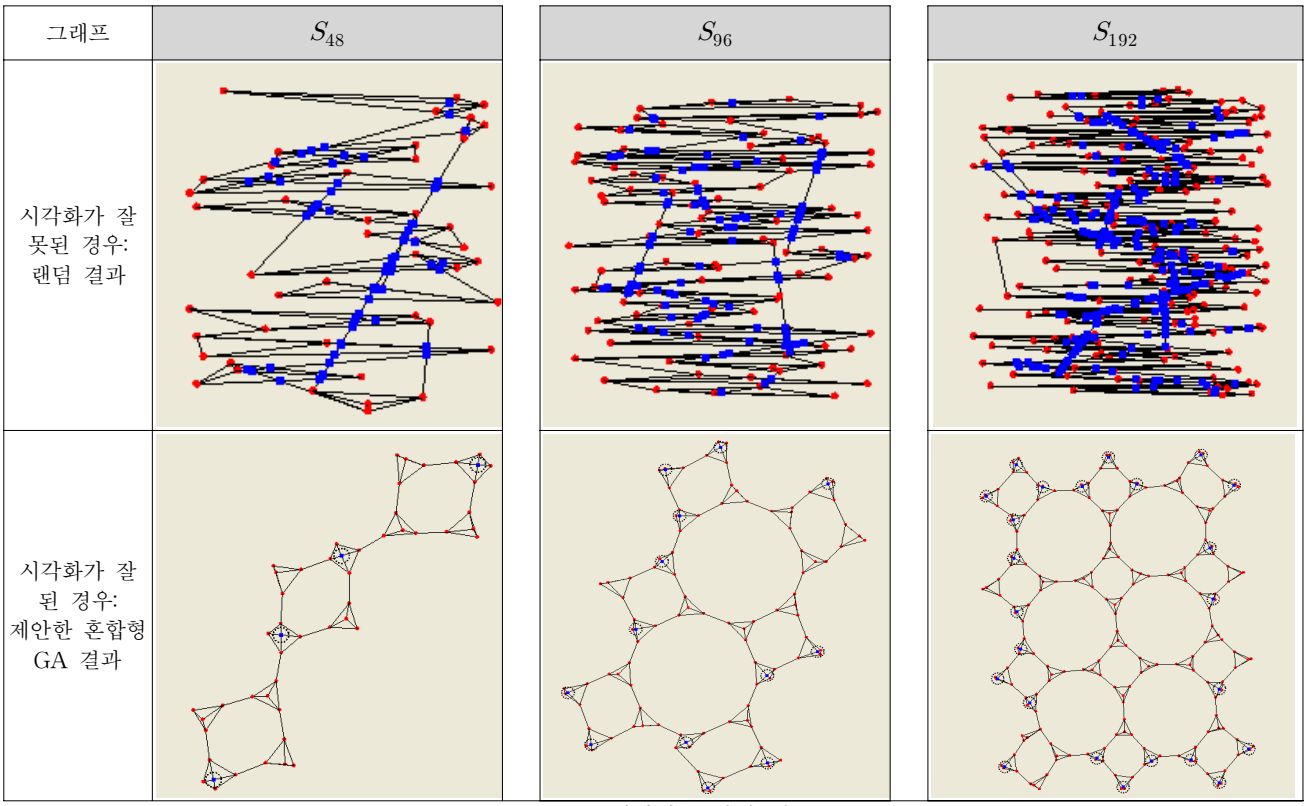


그림 11. 시각화 결과의 예
Fig. 11. Example visualization results

중이 새몬 매핑의 측정값보다 상대적으로 훨씬 커서 GA 연산에서 새몬 매핑의 측정값이 거의 영향을 미치지 못하기 때문인 것으로 보인다.

새몬 매핑을 적용한 경우는 먼저 매핑 시 내부 파라미터인 반복 회수를 정할 때 (반복 회수가 높을수록 새몬 매핑을 더 강하게 수행하게 된다), GA에 새몬 매핑을 지역최적화로 사용하는 혼합형 GA에서는 세대 수가 50이고 새몬 매핑을 각 세대마다 30 회씩 적용하여 새몬 매핑을 총 1,500 회를 실행하기 때문에 이것과 공정한 비교를 하기 위해 반복 회수를 1,500 회로 정하였다. 결과는 앞의 두 경우보다 교차점 수나, 거리 보존의 측면에서 훨씬 좋음을 알 수 있는데, 그 이유는 실험에서 정점 간의 거리가 잘 보존되면 간선의 교차점 수가 비교적 적어지게 되는 그래프를 사용하였고, 표준편차가 0인 것에서 알 수 있듯이 새몬 매핑 측정값이 완전히 수렴했기 때문이다. GA에 새몬 매핑을 지역최적화로 사용한 혼합형 GA는 순수 GA만 사용했을 때 수렴이 너무 늦어지는 단점을 보완하기 위해 제안한 방법이다. GA의 연산을 통해 얻어진 자식 해에 새몬 매핑을 30 회 적용시켜, 만약 새몬 매핑을 적용한 후의 해의 품질이 새몬 매핑을 적용하기 전의 해의 품질보다 나아지는 경우에만 새몬 매핑을 적용한 해를 최종 자식 해로 정하였다. 새몬 매핑만을 적용한 경우와 비교했을 때, 거리 보존 측면에서 약간의 손해가 있지만 거의 차이가 없을 정도로 미미하면서 교차점 개수는 새몬 매핑의 80% 수준으로 줄었음을 확인할 수 있었다.

그림 11은 표 3의 실험 결과 중 랜덤과 혼합형 GA의 최상 해에 대한 시각화 결과를 보여준다. 혼합형 GA의 시각화 결과에서 교차점의 위치를 알아보기 쉽게 하려고 점선 동그라미(O)로 표시하였다. 랜덤으로 그린 그래프는 공통적으로 간선의 교차나 거리의 보존에 대한 고려 없이 임의의 위치에 정점들이 배치되어 있는 것을 확인할 수 있다. 반면에 혼합형 GA 방식은 간선의 교차가 랜덤 그래프에 비해 월등히 적고 정점 간의 거리 보존도 매우 잘 되어서 정점들의 배치가 규칙적으로 일정한 형태를 보이는 것을 확인할 수 있다.

특이할만한 점은 혼합형 GA를 사용하여 간선의 교차가 제거되는 곳이 그래프의 중앙 부분의 교차점에 집중된다는 것이다. 그 이유는 새몬 매핑을 지역 최적화로 사용했을 때 새몬 매핑이 정점 간의 거리를 보존하기 위해 위아래의 양쪽 끝의 교차점이 GA를 통해 제거되었어도 다시 교차하게 만들어 버리는 성질이 있기 때문인 것으로 추측한다. 이러한 성질 때문에 예를 들어 S_{48} 그래프의 경우에 네 개의 교차점 중에 위아래 끝에 있는 교차점은 아무리 실험 회수를 늘여도 제거되지 않았다.

5. 결 론

그래프 시각화는 그래프의 구조를 한 눈에 알아볼 수 있게 해준다는 측면에서 그동안 중요한 이슈가 되어왔다. 본 논문에서는 그래프를 시각화할 때 정점과 간선의 수가 너무 많은 경우에 원래의 그래프 구조를 보존하려고 하면 교차하는 간선이 많아져 정점 간의 관계를 제대로 시각화하기 어려운 문제점을 해결하고자 하였다. 이 논문에서는 정점 간의 거리유지 정도 및 간선의 교차 수를 반영한 평가함수를 사용한 혼합형 유전 알고리즘을 제안하였고, 제안한 방법의 우수성을 실험을 통해 입증했다. 이 방법은 기존의 연구들이 간선 교차만을 줄여 원래의 정점 간의 거리 정보를 손실하는 것에 비해서, 정점 간의 거리를 보존하는 동시에 간선 교차도 최소화한다는 점에서 기존 방법보다 더 직관적인 시각화 방법이라 할 수 있겠다.

본 논문에서 제안한 방법은 더 개선할 여지가 남아있다. 4장의 실험 결과에서 제안한 방법으로는 변두리 부분의 교차점 해소에 상당한 어려움이 있었다. 이 문제의 해결을 포함하여 더욱 개선된 시각화 방법의 고안과 보다 확장된 실험을 통한 시각화 결과 분석을 앞으로의 연구 과제로 남긴다.

참 고 문 헌

[1] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch, “The aesthetics of graph visualization,” *Computational Aesthetics in Graphics, Visualization and Imaging*, pp. 1-8, 2007.

[2] M. Srinivas and L. M. Patnaik, "Genetic algorithms: A survey," *IEEE Computer*, Vol. 27, pp. 17-26, 1994.

[3] G. di Battista, P. Eades, R. Tamassia, and I. G. Tollis, "Algorithms for drawing graphs: An annotated bibliography," *Computational Geometry: Theory and Applications*, Vol. 4, No. 5, pp. 235-282, 1994.

[4] I. Herman, G. Melançon, and M. S. Marshall. “Graph visualization and navigation in information visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 6, pp. 24-43, 2000.

[5] H. C. Purchase, “Which aesthetic has the greatest effect on human understanding?" *In Proceedings of the Symposium on Graph Drawing*, pp. 248-261, 1997.

[6] H. C. Purchase, R. F. Cohen, and M. James, “Validating graph drawing aesthetics," *In Proceedings of the Symposium on Graph Drawing*, pp. 435-446, 1995.

[7] H. C. Purchase, R. F. Cohen, and M. James, “An experimental study of the basis for graph drawing algorithms," *ACM Journal of Experimental Algorithmics*, Vol. 2, No. 4, 1997.

[8] M. R. Garey and D. S. Johnson, “Crossing number is NP-complete," *SIAM Journal of Algebraic and Discrete Methods*, Vol. 4, No. 3, pp. 312-316, 1983.

[9] P. Eades and S. H. Whitesides, “Drawing graphs in two layers," *Theoretical Computer Science*, Vol. 131, No. 2, pp. 361-374, 1994.

[10] W. Tutte, “How to draw a graph," *In Proceedings of the London Mathematical Society*, Vol. 3, No. 13, pp. 743-768, 1963.

[11] M. Juenger and P. Mutzel, “2-layer straightline crossing minimization: Performance of exact and heuristic algorithms," *Journal of Graph Algorithms and Applications*, Vol. 1, pp. 33-59,

1997.
[12] M. R. Laguna, R. Martí, and V. Vals, "Arc crossing minimization in hierarchical digraphs with tabu search," *Computers and Operations Research*, Vol. 24, No. 12, pp. 1165-1186, 1997.
[13] M. Laguna and R. Martí, "GRASP and path re-linking for 2-layer straight line crossing mini-mization," *INFORMS Journal on Computing*, Vol. 11, pp. 44-52, 1999.
[14] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.
[15] J. W. Sammon, Jr, "A non-linear mapping for data structure analysis," *IEEE Transactions on Computers*, Vol. 18, pp. 401-409, 1969.
[16] Y.-H. Kim and B.-R. Moon, "New usage of Sammon's mapping for genetic visualization," *In Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1136-1147, 2003.

저 자 소 개

계주성 (Ju-Sung Kye)
2007 : 광운대학교 컴퓨터소프트웨어학과 (학사)
2009 : 광운대학교 컴퓨터과학과 (석사)

관심분야 : 유전 알고리즘, 데이터베이스
Phon : 02) 940-5217
E-mail : kyejusung@kw.ac.kr

김용혁 (Yong-Hyuk Kim)
1999 : 서울대학교 전산과학 전공 (학사)
2001 : 서울대학교 컴퓨터공학부 (석사)
2005 : 서울대학교 컴퓨터공학부 (박사)
2005～2007 : 서울대학교 반도체공동연구소 연구원
2007～현재 : 광운대 컴퓨터소프트웨어학과 조교수

관심분야 : 최적화, 진화연산, 지식공학
Phone : 02) 940-5212
E-mail : yhdfly@kw.ac.kr

김우생 (Woo-Saeng Kim)
1985 : Univ. of Texas at Austin 전산학 (학사)
1987 : Univ. of Minnesota 전산학(석사)
1991 : Univ. of Minnesota 전산학(박사)
2001 : UC 버클리 대학 교환 교수
1992～현재 : 광운대학교 컴퓨터소프트웨어학과 교수

관심분야 : 멀티미디어, 데이터베이스
Phone : 02) 940-5217
E-mail : kwsrain@kw.ac.kr