한국측량학회지, 제 30권 제 2호, 2012. 4 : pp. 125~134

응집 계층 군집화 기법을 이용한 이종 공간정보의 M:N 대응 클래스 군집 쌍 탐색

Detection of M:N corresponding class group pairs between two spatial datasets with agglomerative hierarchical clustering

허 용[1] · 김정옥[2] · 유기윤[3]
Huh, Yong · Kim, Jungok · Yu, Kiyun

Abstract

In this paper, we propose a method to analyze M:N corresponding relations in semantic matching, especially focusing on feature class matching. Similarities between any class pairs are measured by spatial objects which coexist in the class pairs, and corresponding classes are obtained by clustering with these pairwise similarities. We applied a graph embedding method, which constructs a global configuration of each class in a low-dimensional Euclidean space while preserving the above pairwise similarities, so that the distances between the embedded classes are proportional to the overall degree of similarity on the edge paths in the graph. Thus, the clustering problem could be solved by employing a general clustering algorithm with the embedded coordinates. We applied the proposed method to polygon object layers in a topographic map and land parcel categories in a cadastral map of Suwon area and evaluated the results. F-measures of the detected class pairs were analyzed to validate the results. And some class pairs which would not detected by analysis on nominal class names were detected by the proposed method.

Keywords : M:N corresponding class group pair. Graph embedding, Agglomerative hierarchical clustering, Topographic map, Cadastral map

초 록

본 연구는 두 공간정보의 대응 클래스 군집 쌍 탐색을 중심으로 의미론적 정합과정에서 발생하는 M:N 대응관계를 분석하는 방법을 제안한다. 객체의 공유 관계를 이용하여 클래스의 유사도를 측정하고 높은 유사도를 가지는 클래스들을 군집화함으로써 M:N 대응관계를 탐색하고자 한다. 클래스 사이의 유사도를 그래프 모형으로 표현하고 그래프 임베딩 기법을 적용하여 투영공간에서 클래스 사이의 거리가 클래스 중첩분석에 의한 국지적 유사도에 반비례하도록 개별 클래스들의 투영좌표를 계산하고 군집화를 수행함으로써 계층적 대응 군집 쌍을 탐색할 수 있다. 제안된 방법을 평가하기 위하여 경기도 수원시의 수치지형도와 연속지적도에 적용하여 수치지형도의 면 객체 레이어와 연속지적도의 필지 지목의 대응 군집 쌍을 탐색하였다. 탐색된 대응 클래스 쌍의 F-measure를 측정한 결과 약 0.80에서 0.35 사이의 다양한 값을 얻을 수 있었으며, 클래스 명칭과는 상이한 다양한 대응관계를 얻을 수 있었다.

핵심어 : M:N 대응 클래스 군집 쌍, 그래프 임베딩, 응집 계층 군집화, 수치지형도, 연속지적도

1. 서 론

위치기반 서비스 산업이 성장하고 공간정보 취득 및 처리기술이 발전함에 따라 이종 공간정보의 통합에 대한 관심이 증가하고 있다. 그 결과 전통적으로 공간정보를 생산하던 공공기관뿐만 아니라 인터넷 포털 및 모바일 산업을 중심으로 민간분야의 CP(content provider) 업체들은 기구축 공간정보를 가공하고 필요한 부가 정보를 추가함으로써

1) The Hong Kong Polytechnic Univ. Dept. of LSGI Professional Research Fellow (E-mail : huhyong78@gmail.com)
2) 정회원 · 서울대학교 공학연구소 선임연구원(E-mail : geostar1@snu.ac.kr)
3) 교신저자 · 정회원 · 서울대학교 공과대학 건설환경공학부 부교수(E-mail : kiyun@snu.ac.kr)

새로운 공간정보를 생산하고 있다. 공간정보는 사용자 요구에 따라 현실세계의 정보를 선별하여 생산하기 때문에 이종 공간정보를 통합함으로써 개별 공간정보의 취득 과정에서 누락되는 현실세계의 정보를 교환 또는 보완하는 것은 물론, 정확도와 최신성과 같은 품질을 개선하는 것이 가능하다.

공간정보를 통합하기 위해서는 객체와 속성의 대응관계가 필요하기 때문에 객체 정합(object matching)과 의미론적 정합(sematic matching)을 수행하게 된다. 일반적인 정합 과정은 대응관계를 탐색할 정보의 특성을 고려하여 유사도 측정 방법을 결정하고 가장 높은 유사도를 가지는 대응쌍을 탐색한다. 하지만 후보 정합 쌍의 쌍대 비교를 이용한 유사도 측정은 기본적으로 1:1 대응관계를 가정하기 때문에 M:N 대응관계를 탐색하기 위해서는 새로운 방법이 필요하다.

객체 정합의 경우 M:N 대응관계 문제를 해결하기 위하여 다양한 연구가 수행(Bel Hajd Ali, 2000; Kiedler, 2007; 허용 등, 2009; Huh 등, 2011)된 것에 비하여 의미론적 정합은 이종 데이터베이스 사이의 정확한 정보검색을 위한 질의 또는 질의에 필요한 속성명을 상호 생성하고 해석할 수 있는 wrapper 또는 mediator 개발을 목적으로 수행되었고, 대부분 1:1 대응관계를 중심으로 연구되었다(Euzenat and Shvaiki, 2007). 하지만 풍부한 정보를 도출하기 위해서는 대응 객체의 탐색과 마찬가지로 M:N 대응관계를 고려해야 한다. 예를 들어 그림 1과 같이 수치지형도와 연속지적도의 클래스 대응관계 탐색에서 1:1 대응관계만 고려한다면 동일한 지형지물을 표현하는 클래스 쌍을 분석할 수 없다. 이는 클래스로 표현하고자 하는 지형지물의 구분 및 정의가 두 지도에서 상이하기 때문이다. 예를 들어 복개도로라는 지형지물은 수치지형도에서는 도로경계(A001)과 암거(C007)에 모두 포함되는 반면 연속지적도에서는 구거에 대응된다. 개인도로의 경우 공공도로와 함께 수치지형도에서는 도로경계(A001)에 대응되지만 연속지적도에서는 건물과 함께 대지에 대응된다. 따라서 올바른 대응 클래스 쌍을 분석하기 위해서는 현실세계의 공공도로, 개인도로, 복개도로를 포괄하는 상위 클래스(super-class)를 도출하고 수치지형도의 {도로경계, 암거}와 연속지적도의 {도로, 구거}로 구성되는 2:2 대응 클래스 군집 쌍을 탐색해야 한다.

본 연구는 두 공간정보의 대응 클래스 군집 쌍 탐색을 중심으로 의미론적 정합과정에서 발생하는 M:N 대응관계를

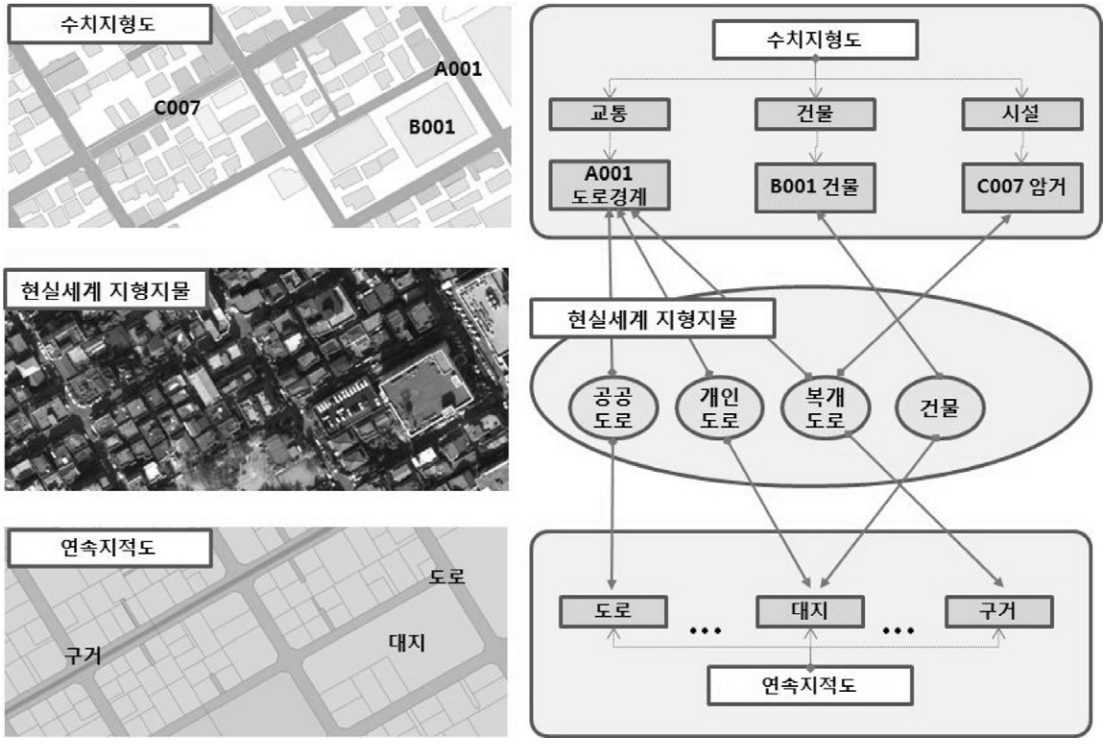


그림 1. 이종 공간정보의 속성정보의 M:N 대응관계 발생 사례
수치지형도와 연속지적도의 클래스 대응관계

탐색하는 방법을 제안한다. 대응 쌍을 탐색하기 위한 클래스 유사도는 해당 클래스에 포함되는 객체의 공유 수준을 이용하여 측정할 수 있다(Duckham 등, 2005; Kieler, 2007). 본 연구에서는 클래스를 노드로, 임의의 두 클래스 사이의 유사도를 에지 가중치로 표현하는 이분 그래프(bipartite graph) 모형에서 노드 군집화를 수행함으로써 M:N 대응관계를 탐색하고자 한다. 하지만 그래프 모형은 수학적 연산이 복잡하기 때문에 그래프 임베딩(graph embedding) 기법을 적용하여 낮은 차원의 벡터 공간에 노드를 투영한 후, 투영된 노드의 벡터 좌표를 이용하여 군집화를 수행하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 대응 클래스 쌍을 탐색하기 위한 선행연구를 분석한다. 3장에서는 그래프 모형으로 표현된 클래스 유사도 정보를 벡터 공간에 투영하기 위한 그래프 임베딩 기법을 설명한다. 4장에서는 클래스 유사도의 측정 방법과 함께 투영된 공간에서의 군집화 기법에 대하여 설명한다. 5장에서는 제안된 기법을 수치지형도와 연속지적도에 적용한 결과를 분석하고, 마지막으로 6장에서 결론을 제시한다.

2. 선행연구 분석

공간정보의 클래스는 지형지물을 공간객체로 표현하기 위한 추상화 과정에서 결정된다(Uitermark 등, 1999). 여기서 추상화 과정이란 그림 1의 공공도로, 개인도로, 복개도로, 건물과 같이 일반적인 지식이나 개념에 기초하여 지형지물을 분류한 뒤, 공간정보의 활용목적에 적합한 개념적 구조 및 세밀도(conceptual hierarchy and granularity)에 따라 제구성하는 것을 의미한다(Uitermark 등, 1999). 따라서 명칭은 동일하지만 의미하는 바가 상이하거나 반대로 명칭은 상이하지만 의미하는 바가 동일한 문제를 해결하기 위해서는 활용분야의 개념이나 어휘에 대한 분석이 필요하다. 일반적으로 이 분석은 충분한 배경지식을 가지고 있는 전문가에 의하여 수행되었다(황보택근, 이기정, 2006; Kokla, 2006).

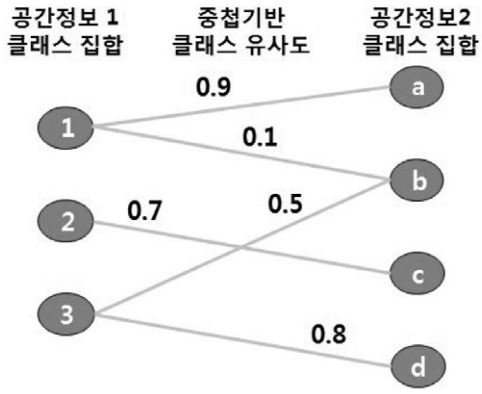
하지만 전문가 분석은 전문자의 사전지식에 의존하기 때문에 새로운 유형의 공간정보인 경우 상대적으로 많은 분석 시간을 요구한다. 이 문제를 해결하기 위하여 객체 기반의 대응 클래스 탐색 연구가 수행되고 있다. Uitermark 등(1999)은 클래스에 포함되는 객체 집합의 중첩분석을 수행한 뒤, 클래스 대응관계를 분석하였다. Duckham 등(2005)은 중첩분석을 수행하여 얻어진 클래스의 대응관계를 형식개념분석(formal concept analysis)의 개념 격자(concept lattice) 모형에 적용하여 계층적 대응관계를 탐색하였다. Kieler(2007)는 클래스 사이의 중첩 면적비를 이용하여 독일의 ATKIS(authoritative topographic cartographic information system) 자료를 중심으로 차량 항법용 공간정보인 GDF(geographic data files)와 지적관리를 위한 ALK(automated real-estate map)의 클래스 대응관계를 분석하였다. Yi 등(2007)은 클래스를 노드로, 두 클래스의 관계를 에지로 표현하는 ARG(attributed relational graph)를 유도한 뒤, 그래프 정합을 수행함으로써 대응 클래스를 탐색하였다. Parundekar 등(2010)은 온톨로지 정합 기법을 적용하여 두 공간정보 속성 집합의 계층적 대응 관계를 탐색하였다. INSPIRE(infrastructure for spatial information in the European community)에서는 HUMBOLDT 프로젝트를 수행하여 공간정보 융합(harmonization) 플랫폼을 개발하고 클래스 대응관계를 포함한 의미론적(semantic) 정합을 수행할 수 있는 방법론과 HALE(HUMBOLDT Alignment Editor)를 공개하였다(Fichtinger 등, 2011).

앞에서 언급한 선행연구에서 M:N 대응관계가 발생한 경우 클래스 유사도에 임계값을 적용하여 대응 여부를 결정하고 이렇게 연결된 클래스들을 하나의 군집으로 탐색하였다(Uitermark 등,1999; Duckham 등, 2005; Kieler, 2007). 하지만 임계값에 따라 대응 여부가 불연속적으로 결정되며, 1:1, 1:N, M:N 대응 관계별로 서로 다른 임계값을 어떻게 결정할 것인가에 대한 기준을 제시하지 못하였다.

3. k차원 그래프 임베딩

본 연구에서는 클래스를 노드, 클래스 쌍의 유사도를 에지 가중치로 가지는 그래프 모형을 이용하여 대응 클래스 군집 쌍을 탐색한다. 두 공간정보의 클래스 집합이 각각 V_1과 V_2로 주어졌을 때, V_1과 V_2의 합집합이 그래프의 노드 집합 $V=\{v_i|v_i\in V_1\cup V_2\}$가 된다. 에지 집합은 V_1과 V_2사이의 클래스 중첩 여부를 이용, $E=\{(v_i,v_j)|\ v_i\cap v_j\neq\varnothing, v_i\in V_1, v_j\in V_2\}$와 같이 얻어진다. 만약 에지에 가중치가 부여된다면 노드 집합의 노드 개수 $|V|$만큼의 행과 열을 가지는 에지 가중치 행렬 $W=\{w_{ij}|v_i\in V, v_j\in V\}$로 표현한다. 만약 3개와 4개의 클래스를 가지는 두 공간정보에서 클래스 유사도가 그림 2(a)와 같다면 그림 2(b)와 같이 노드 집합, 에지 집합 그리고 에지 가중치 행렬로 표현할 수 있다.

행렬 W의 i번째 행은 노드 v_i와 나머지 노드 사이의 유



(a)

$V_1 = \{1,2,3\}$, $V_2 = \{a,b,c,d\}$

$E = \{(1,a),(1,b),(2,c),(3,b),(3,d)\}$

$w_{i,j}$	1	2	3	a	b	c	d
1	-	-	-	0.9	0.1	-	-
2	-	-	-	-	-	0.7	-
3	-	-	-	-	0.5	-	0.8
a	0.9	-	-	-	-	-	-
b	0.1	-	0.5	-	-	-	-
c	-	0.7	-	-	-	-	-
d	-	-	0.8	-	-	-	-

(b)

그림 2. 두 공간정보의 클래스 유사도 측정결과(a)의 그래프 모형 표현(b)

사도를 $1\times|V|$의 크기를 가지는 특징벡터로 표현한다. 그래프 임베딩은 각 노드의 $|V|$차원 특징벡터를 k차원 특징벡터로 압축함으로써 행렬 W를 $|V|\times k$크기의 행렬로 압축한다. 이 과정에서 유사도가 높은 노드, 즉 높은 에지 가중치로 연결된 노드를 새로운 벡터공간에서 인접한 좌표에 위치하도록 최적화함으로서 행렬 W의 정보를 최대한 유지한다. 따라서 투영된 노드의 거리를 측정하면 에지 가중치로 표현된 노드 사이의 유사도를 측정할 수 있다.

그래프 임베딩을 이용하여 행렬 W를 1차원에 투영하기 위해서는 식(1)과 같은 목적함수를 최소화시키는 투영 좌표 $\{x_i | i = 1, \cdots, |V|\}$를 탐색해야 한다(Pothen 등, 1990; Yan 등, 2007). 식(1)을 최소화하면 높은 에지 가중치로 연결된 노드일수록 투영 좌표의 차이는 작아지며, 반대로 낮은 에지 가중치로 연결된 노드일수록 투영 좌표의 차이는 커지게 된다.

$$F(x) = \sum_{i,j}(x_i - x_j)^2 w_{ij} \tag{1}$$

여기서, x_i : 노드 v_i의 투영 좌표

식(1)의 우항은 식(2)와 같이 라플라시안 행렬 L과 벡터 x를 이용하여 표현하는 것이 가능하다. 따라서 $\underset{x}{argmin}\ (x^T L\, x)$의 해를 구함으로써 식(1)의 최적화 문제로 해결할 수 있다(Pothen 등, 1990).

$$F(x) = 2\,[x_1, x_2, \cdots, x_{|V|}] \begin{bmatrix} d_{11} & -w_{12} \cdots & -w_{1|V|} \\ -w_{21} & d_{22} & \\ \vdots & \ddots & \vdots \\ -w_{|V|1} & \cdots & d_{|V||V|} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{|V|} \end{bmatrix} \tag{2}$$

$$= 2x^T(D - W)x$$

$$= 2x^T L\ x$$

여기서, d_{ii} : 노드 v_i에 연결된 에지 가중치 합,

$$\sum_{j=1,\, j\neq i}^{|V|} w_{ij}$$

D : d_{ii}를 대각성분으로 가지는 행렬,

$diag(d_{11}, \cdots, d_{|V||V|})$

L : 라플라시안 그래프 행렬, $D - W$

$\underset{x}{argmin}\ (x^T L\, x)$의 해를 얻기 위해서는 두 가지 제약조건을 고려해야 한다. 벡터 x가 $(0, \cdots, 0)^T$인 경우 에지 가중치 $w_{i,j}$는 0 이상이므로 식 (1)의 목적함수 $F(x)$는 0값을 가지게 되어 극소값을 가지게 된다. 또한 0에 가까운 상수 c를 곱한 cx의 경우에도 x의 성분에 무관하게 목적함수는 0에 가까운 값을 가지게 된다. 이 문제를 해결하기 위하여 $x^T x = 1$의 제약조건을 추가한 뒤, 식(3)과 같이 라그랑즈 승수법을 이용하여 최적해 $\tilde{x}$를 얻는다(Hendrickson, 2007). 에지 가중치 행렬 W가 대칭이면 라플라시안 행렬 L 역시 대칭이다. 따라서 식(4)를 식(5)와 같이 전개할 수 있으며, 특이벡터문제의 해로 최적해를 얻을 수 있다.

$$F(x, \lambda) = \langle Lx, x\rangle - \lambda(\langle x, x\rangle - 1) \tag{3}$$

$$\frac{\partial F(x,\lambda)}{\partial x} = x^T(L + L^T) - 2\lambda x^T \tag{4}$$

$$L\tilde{x} = \lambda\,\tilde{x} \tag{5}$$

여기서, $\langle\ \ \rangle$: 벡터 내적 연산자, $\langle x, y\rangle = x^T y$

하지만 식(5)에서 $\tilde{x}$는 특이벡터의 개수만큼 존재하게 된다. Fiedler(1975)는 0이 아닌 최소 특이값에 대응되는 벡터를 이용함으로써 최적해를 얻을 수 있다는 것을 증명하였다. 이후 Sameh 등(1982)은 k 차원 그래프 임베딩으로 최적화 문제로 확장할 수 있는 식(6)을 증명하였다. 따라서 k 차원에서 그래프 노드의 투영좌표를 얻기 위해서는 0이 아닌 k개의 최소 특이값 $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_k$에 대응되는 특이벡터의 성분들을 이용하면 된다.

$$\min trace(X^T L X) = \sum_{i=1}^{k} \lambda_i \tag{6}$$

여기서, X : $|V| \times k$의 크기를 가지는

$[\tilde{x}^{(1)}, \ldots, \tilde{x}^{(k)}]$ 행렬

$\tilde{x}^{(i)}$: λ_i에 대응되는 $Lx = \lambda x$의 특이벡터,

단 $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_k$

이렇게 얻어진 k개의 벡터는 $\tilde{x}^{(i)T} \tilde{x}^{(i)} = 1$의 제약조건에 의하여 정규화된다. 식(6)의 관계를 고려하였을 때 작은 특이값에 대응되는 특이벡터일수록 목적함수의 최소화에 적합하다. 따라서 본 연구에서는 식(7)과 같이 특이값의 제곱근 역수를 가중치로 적용하여 각 차원의 가중치를 투영좌표에 반영한다(Trosset 등, 2010).

$$X' = \left[\frac{\tilde{x}^{(1)}}{\sqrt{\lambda_1}}, \ldots, \frac{\tilde{x}^{(k)}}{\sqrt{\lambda_k}} \right] \tag{7}$$

4. 제안된 M:N 대응 클래스 군집 쌍 탐색

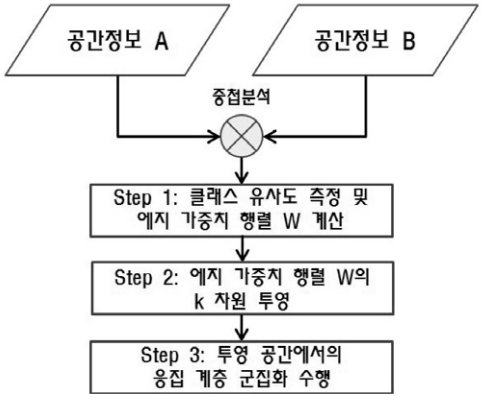


그림 3. 제안된 M:N 대응 클래스 군집 쌍 탐색 순서

본 연구에서 제안하는 M:N 대응 클래스 군집 쌍 탐색 순서는 그림 3과 같이 구성되며, 세부적인 내용은 아래와 같다.

4.1 클래스 유사도 행렬 S 및 에지 가중치 행렬 W계산

M:N 대응 클래스 군집 쌍의 탐색을 위해서는 두 공간정보 클래스 사이의 유사도를 측정해야 한다. 정확한 객체기반 유사도를 측정하기 위해서는 객체 정합을 이용한 대응 객체 쌍을 탐색하는 것이 필요하다. 하지만 객체 정합은 본 연구의 범위를 초과하므로 다음과 같이 클래스에 포함된 전체 객체의 면적을 이용하여 중첩분석을 수행하고 조건부 확률 개념을 이용하여 식(8)과 같이 유사도를 측정한다(Bel Hadj Ali, 2001; Kieler, 2007). $\Pr(a_i|b_j)$은 공간정보 B의 클래스 b_j의 공간객체가 공간정보 A의 클래스 a_i의 공간객체에 중첩될 확률로 b_j클래스에 포함된 전체 공간객체의 면적과 중첩면적을 이용하여 측정한다. 하지만 식(8)의 유사도는 분모에 따라 $\Pr(a_i|b_j)$와 $\Pr(b_j|a_i)$가 상이하기 때문에 $S(a_i, b_j)$와 $S(b_j, a_i)$가 동일하지 않은 문제를 가진다.

$$S(a_i, b_j) = \Pr(a_i|b_j) = \frac{Area(a_i \cap b_j)}{Area(b_j)} \tag{8}$$

여기서, x_k : 공간정보 X의 k번째 클래스에 포함된 전체 공간객체의 면적

앞의 식 (4)를 식(5)와 같이 전개하기 위해서는 유사도의 비대칭성 문제를 해결해야 하며, 본 연구에서는 $S(a_i, b_j)$와 $S(b_j, a_i)$의 평균을 두 클래스의 유사도로 이용한다. 따라서 두 공간정보 A와 B의 클래스 개수의 합이 일 N때 최종 에지 가중치 행렬 W는 식(9)와 같이 결정된다.

$$W = \frac{1}{2}(S + S^T) \tag{9}$$

여기서, S : 식(8)을 이용하여 측정된 $N \times N$크기의 클래스 유사도 행렬

4.2 에지 가중치 행렬 W의 k차원 투영

3절에서 설명한 그래프 임베딩을 수행하기 위해서는 투영될 공간의 차원 k이 결정되어야 한다. Dhillon (2001)은 최적 k값을 결정하기 위하여 식(10)과 같이 두 공간정보의 클래스 개수 중 작은 클래스 개수에 로그값을 취하고 올림

함수를 이용하여 로그값보다 큰 가장 작은 정수를 제안하였다. 이것은 군집화될 클래스 군집 쌍의 개수는 공간정보의 클래스 개수를 초과할 수 없다는 점과 임의의 차원에서의 좌표 $x^{(i)}$를 이용하면 최소 2개 이상의 군집으로 분할할 수 있다는 가정에 근거한다.

$$k = \lceil \log_2(\min(n_1, n_2)) \rceil \quad (10)$$

여기서, $\lceil \quad \rceil$: 올림 함수

에지 가중치 행렬 W와 투영될 차원의 개수 k가 결정되면 식(5)와 같이 $Lx = \lambda x$에서 $0 = \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_k$의 조건을 만족하는 가장 작은 k개의 특이값 $\lambda_1, \cdots, \lambda_k$와 이에 대응되는 특이벡터를 계산하고 식(7)과 같이 투영좌표를 계산한다.

4.3 투영 공간에서의 응집 계층 군집화 수행

일반적인 군집화와 달리 본 연구에서 군집을 구성하는 클래스의 개수는 1:1 대응관계의 경우 2개 그리고 M:N 대응관계수의 경우 M+N과 같이 상대적으로 매우 작다. 따라서 정규분포와 같은 군집의 분포를 가정하는 알고리즘을 적용할 경우 왜곡된 결과를 얻을 수 있다. 이 문제를 해결하기 위하여 군집의 분포 및 구조에 관한 가정에 큰 영향을 받지 않는 응집 계층 군집화(agglomerative hierarchical clustering) 기법을 이용한다. 이 기법은 개별 클래스를 초기 군집으로 가정하고 군집 사이의 거리를 측정한다. 이후 가장 인접한 두 클래스를 하나의 클래스로 군집화한다. 이후 클래스 군집 사이의 거리를 다시 측정하고 가장 인접한 클래스 두 개를 하나로 군집화하는 과정을 전체 클래스가 하나의 군집이 될 때까지 반복한다(오일석, 2008). 이 과정은 트리 형태의 덴드로그램으로 표현되며, 상향식 탐색을 수행함으로써 대응 클래스의 후보 군집 쌍을 탐색한다.

5. 실험 및 평가

5.1 실험 대상 자료

제안된 방법을 평가하기 위하여 표 1과 같이 수원 지역의 KLIS 연속지적도와 수치지형도 2.0에 적용하였다. 연속지적도의 경우 필기기반의 면 객체로 구성되어 있지만 수치지형도의 경우 점, 선 그리고 면 객체로 구성되어 있으므로 면 객체로 구성된 10개의 레이어를 이용하였다. 그런데 연속지적도와는 달리 수치지형도에는 면 객체로 표현되지 않는 빈 공간이 존재하는데 이러한 공간에 대응되는 '기타' 라는 새로운 수치지형도 레이어를 추가하였다. 이렇게 추가된 수치지형도의 11개 클래스와 본 연구 대상지역에는 존재하지 않는 목장용지, 염전 등을 제외한 24개 지목과 아직 지목이 부여되지 않는 필지들을 포함한 연속지적도의 25개 클래스의 대응 클래스 군집 쌍을 분석한다.

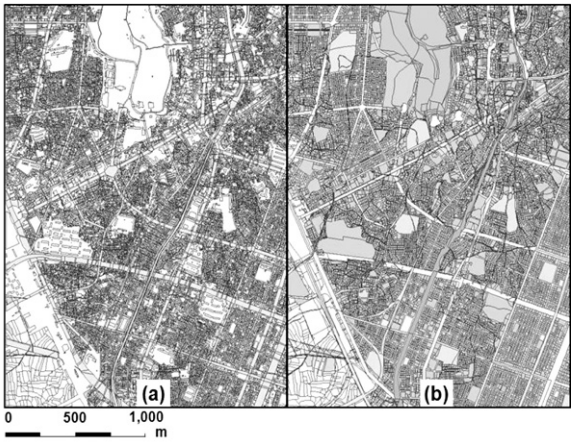


그림 4. 실험 대상지역 (a) 수치지형도 (b) KLIS 연속지적도

표 1. 실험 대상 자료

	KLIS 연속지적도	수치지형도 2.0
제작기관	대한지적공사	국토지리정보원
파일형식	shp	ngi
대상지역	경기도 수원시 팔달구 지역(2km×3km)	
제작시점	2008.06	2006.12
축척	1:1,200	1:5,000
분석 대상 클래스 개수	11개	25개

5.2 실험 결과

두 지도의 클래스 면적 및 중첩면적을 식(8)에 적용하여 유사도를 측정하여 식(9)와 같은 가중치 행렬 W를 계산하였다. 표 2는 에지 가중치 행렬 W에서 식(11)과 같은 관계를 가지는 좌상단 부분행렬 w이다. 이 때 에지 가중치 w_{ij}는 $0.5*\Pr(a_i|b_j)+0.5*\Pr(b_j|a_i)$와 같이 측정된다. 따라서 $\{a_1\}:\{b_1,b_2,b_3\}$과 같은 1:3 관계가 발생할 경우 $\Pr(a_1|b_j)$은 1에 가까운 값을 가지므로 측정된 클래스들의 에지 가중치 w_{1j}는 모두 0.5 이상이 된다. 따라서 1:N과 같은 전체-부분 관계에서 전체에 해당하는 클래스의 에지 가중치 합은 다른 클래스들에 비하여 높은 값을 가지게 된다.

$$W=\begin{bmatrix}0 & w\\ w^T & 0\end{bmatrix} \tag{11}$$

두 지도의 클래스 개수는 각각 11개와 25개로 식(10)를 이용하면 투영될 차원의 개수가 4개로 결정된다. 표 2의 에지 가중치를 그래프 임베딩을 이용하여 투영하였을 때 차원별 좌표는 그림 5의 (a)와 (b)와 같다. 앞에서 설명한 바와 같이 투영좌표를 이용하여 응집 계층 군집화를 수행하면 그림 5(c)와 같은 덴드로그램을 얻게 된다.

그림 5(c)의 덴드로그램에서 링크를 분석함으로써 {호수/저수지}:{유지}나 {지류계}:{답}과 같은 1:1 대응 쌍은 물론, {지류계}:{답, 기타}이나 {도로경계}:{도로, 구거}와 같은 1:2 대응 쌍을 확인할 수 있다. 반면 과수원, 학교용지, 주차장, 임야, 묘지 등은 수치지형도의 '기타' 레이어에 대응된다. 이는 수치지형도에서 공간 객체로 표현되지 않는 빈 공간에 해당 지목을 가진 필지들이 분포하기 때문으로 판단된다.

표 2. 식(9)를 이용하여 측정된 수치지형도와 연속지적도의 에지 가중치 행렬 W의 좌상단 부분행렬 w (기타 : 수치지형도에서 공간객체가 없는 공간)

	미 정	전	답	과수원	임 야	대 지	공장용지	학교용지	주차장	주유소용지	창고용지	도 로	철도용지
도로경계(A001)	0.3730	0.1512	0.0461	-	0.0573	0.0647	0.0336	0.0041	0.0130	0.0025	0.0073	0.8014	0.1064
철도경계(A016)	0.0007	0.0053	0.0493	-	-	0.0166	0.0717	-	-	-	-	0.0063	0.6490
건물(B001)	0.0277	0.0996	0.0148	0.1959	0.0154	0.7602	0.1116	0.2211	0.1943	0.3029	0.4932	0.0143	0.1365
탱크(C018)	-	-	-	-	-	0.1101	0.3723	-	-	-	-	-	0.0215
계단(C039)	-	-	-	-	0.0074	0.3767	-	0.0517	-	-	-	0.0035	-
주유소(C042)	-	-	0.0029	-	-	0.2813	-	0.0007	-	0.3573	-	0.0248	0.0017
주차장(C043)	-	0.0007	-	-	-	0.4636	-	-	-	-	-	0.0086	-
지류계(D002)	-	0.0792	0.4136	-	0.0041	0.0029	-	0.0003	0.0389	-	-	0.0009	0.0135
하천경계(E001)	0.1221	-	0.0061	-	-	0.0004	-	-	-	-	-	0.0207	0.0286
호수/저수지(E005)	-	-	-	-	-	0.0417	-	-	-	-	-	0.0230	-
기 타	0.0402	0.2048	0.0458	0.3044	0.5141	0.4499	0.3014	0.3367	0.2550	-	-	0.0965	0.0125

	제 방	하 천	구 거	유 지	수도용지	공 원	체육용지	유원지	종교용지	사적지	묘 지	기 타
도로경계(A001)	-	0.1933	0.3337	0.0265	-	0.0121	0.0025	0.0273	0.0069	0.0548	0.0493	0.0145
철도경계(A016)	-	0.0015	0.0102	-	-	-	-	-	-	-	-	0.0001
건물(B001)	0.0023	0.0153	0.0294	0.0022	0.0064	0.0278	0.0461	-	0.4331	0.1099	0.0116	0.0573
탱크(C018)	-	-	-	-	-	-	-	-	-	-	-	-
계단(C039)	-	-	-	-	-	0.0833	-	-	0.0006	0.0022	-	-
주유소(C042)	-	0.0009	0.0229	-	-	-	-	-	0.0009	-	-	0.0078
주차장(C043)	-	0.0099	0.0202	-	-	-	-	-	-	-	-	-
지류계(D002)	-	0.0172	0.0103	-	-	-	-	-	0.0029	-	0.0068	0.3192
하천경계(E001)	0.0164	0.5694	0.0036	0.0179	-	-	-	-	-	0.0114	-	-
호수/저수지(E005)	-	-	0.0769	0.6389	-	-	-	-	-	-	-	-
기 타	0.4823	0.1117	0.1032	0.2132	0.4949	0.4464	0.4579	0.4730	0.0619	0.3284	0.4370	0.1375

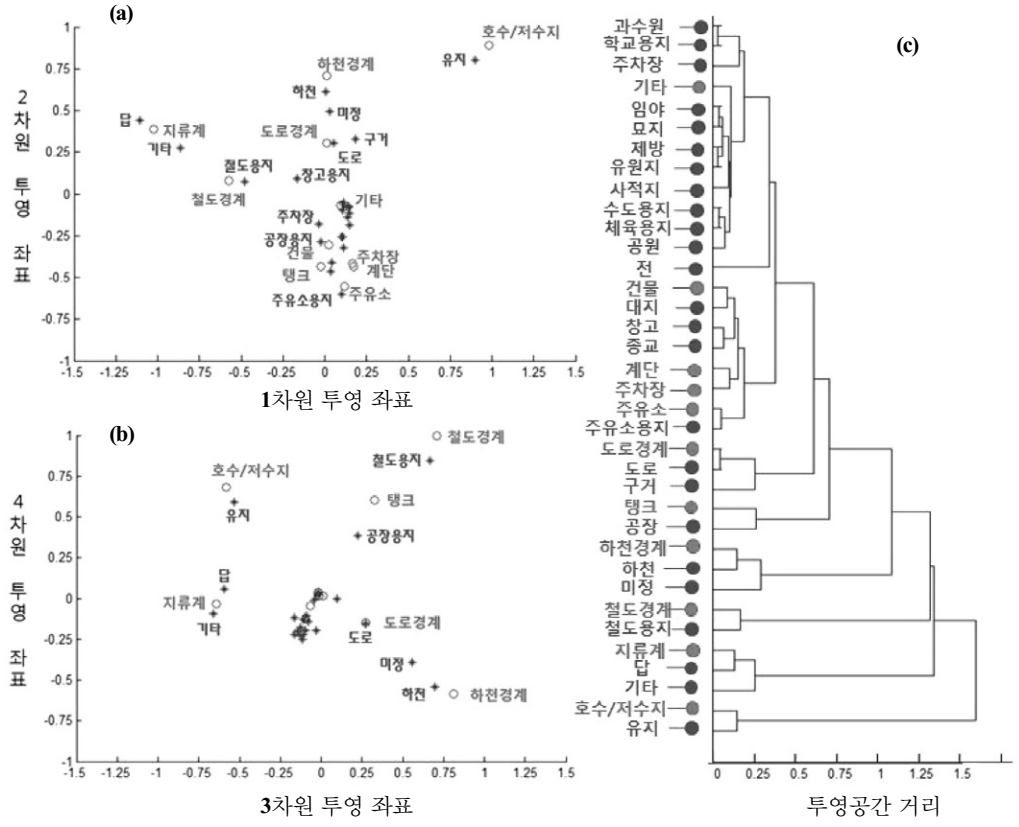


그림 5. 그래프 임베딩을 수행한 4차원 투영공간에서 클래스의 좌표(a) (b)와 투영 공간에서의 거리를 이용한 응집 계층 군집화를 수행한 덴드로그램(c)

군집화 결과를 정량적으로 분석하기 위하여 덴드로그램에서 k 번째 군집을 구성하는 수치지형도의 레이어 집합 $\{a_1^{(k)}, \cdots, a_n^{(k)}\}$ 와 연속지적도의 지목 집합 $\{b_1^{(k)}, \cdots, b_m^{(k)}\}$ 이 있을 때, 각 군집의 객체들을 병합한 뒤, 중첩면적을 이용하여 식(12)와 같이 F-measure를 측정하였다(Euzenat 등, 2007).

$$F-measure = 2\frac{P_a^{(k)} \times P_b^{(k)}}{P_a^{(k)} + P_b^{(k)}} \tag{12}$$

여기서, $P_a^{(k)}: \dfrac{Area\{a_1^{(k)}, \cdots, a_n^{(k)}\} \cap Area\{b_1^{(k)}, \cdots, b_m^{(k)}\}}{Area\{a_1^{(k)}, \cdots, a_n^{(k)}\}}$

$$P_b^{(k)}: \frac{Area\{a_1^{(k)}, \cdots, a_n^{(k)}\} \cap Area\{b_1^{(k)}, \cdots, b_m^{(k)}\}}{Area\{b_1^{(k)}, \cdots, b_m^{(k)}\}}$$

{도로경계}:{도로}의 F-measure는 0.7981로 수치지형도의 도로 레이어와 연속지적도의 도로 지목은 매우 높은 유사도를 가지고 있음을 확인할 수 있으며, {도로경계}:{도로, 구거}는 이보다 높은 0.8077의 유사도를 가지고 있다. 따라서 수치지형도의 A001에 포함되는 객체의 대응 쌍을 탐색하기 위해서는 연속지적도에서 도로 및 구거의 지목을 가지는 필지를 동시에 고려하는 것이 더 타당함을 알 수 있다. 도로와 마찬가지로 {철도경계}:{철도용지}, {지류계}:{답}, {하천경계}:{하천} 그리고 {호수/저수지}:{유지}와 같이 명칭이 유사한 경우 대부분 높은 F-measure가 측정되었다. 하지만 {C042}:{주유소용지}와 같이 유사한 명칭을 가지고 있지만 낮은 유사도를 가지는 경우도 발생한다. 이는 주유소를 공간 객체로 표현하는 기준이 두 지도에서 상이하기 때문이다. 수치지형도에서는 주유소의 건물을 건물 레이어의 객체로 표현하고 나머지 면적을 주유소 레이어의 객체로 표현하는 반면, 연속지적도에서는 두 객체를 모두 포함하는 필지의 개념으로 주유소용지 객체를 생성하기 때문이다. 반면 연속지적도에는 건물에 해당하는 지목은 존재하지 않지만 대부분의 건물이 대지 필지 위에

표 3. 제안된 탐색 기법으로 탐색된 주요 1:N 및 M:N 대응 클래스 쌍

수치지형도 레이어	연속지적도 지목	F-measure
도로경계	도로	0.7981
철도경계	철도	0.6004
주유소	주유소용지	0.3541
건물	대지	0.7369
건물	대지, 창고용지, 종교용지	0.4415
지류계	답	0.6132
하천경계	하천	0.4934
호수/저수지	유지	0.6116
건물, 계단, 주차장	대지, 창고용지, 종교용지	0.7437
건물, 계단, 주차장, 주유소	대지, 주유소용지, 창고용지, 종교용지	0.7458
지류계	답, 기타	0.7609
도로경계	도로, 구거	0.8077
건물, 계단, 주유소, 주차장, 기타	전, 과수원, 임야, 대지, 학교용지, 주차장, 주유소용지, 창고용지, 제방, 수도용지, 공원, 체육용지, 유원지, 종교용지, 사적지, 묘지	0.9363

존재하기 때문에 {건물}:{대지}의 F-measure가 0.7369로 높게 측정되었다. {건물, 계단, 주유소, 주차장, 기타}:{전, 과수원, 임야, 대지, 학교용지, 주차장, 주유소용지, 창고용지, 제방, 수도용지, 공원, 체육용지, 유원지, 종교용지, 사적지, 묘지}와 같이 군집의 규모가 커질수록 F-measure는 매우 높아지지만 유의미한 클래스 정보를 제공하지 않기 때문에 이후의 대응 클래스 군집 쌍의 분석은 수행하지 않았다.

6. 결론 및 향후 연구

본 연구는 두 공간정보의 대응 클래스 군집 쌍 탐색을 중심으로 M:N 대응관계를 탐색하는 새로운 방법을 제안하였다. 클래스에 포함되는 공간객체 집합의 중첩면적을 이용하여 클래스의 유사도를 측정하고 그래프 임베딩 기법을 적용하여 낮은 차원의 벡터 공간에 클래스를 투영하였다. 이 때 높은 유사도를 가지는 클래스는 투영 공간에서 인접한 위치에 분포하기 때문에 거리 기반의 군집화를 수행함으로써 높은 유사도를 가지는 클래스들을 효과적으로 탐색할 수 있다. 제안된 방법을 평가하기 위하여 경시도 수원 지역의 수치지형도와 연속지적도에 적용하였다. 수치지형도의 11개 레이어와 연속지적도의 25개 지목을 대상으로 대응 클래스 군집 쌍을 탐색한 결과 명칭이 유사한 경우 대부분 1:1 대응관계를 가지고 있었다. 하지만 주유소의 경우처럼 두 지도에서 공간 객체로 표현하는 기준이 다를 경우 명칭만을 이용하는 탐색은 잘못된 결과를 가져올 수 있음을 확인할 수 있었다. 또한 도로의 경우 수치지형도의 도로경계 레이어의 객체는 단순히 도로 지목을 가지는 객체들보다 도로와 구거의 지목을 가지는 필지들을 동시에 고려하였을 때 더 높은 유사도를 가짐을 알 수 있었다.

본 연구에서 대응 클래스 쌍을 탐색하기 위하여 분석 대상 공간정보의 중첩분석만을 이용한다. 따라서 공간정보 구축에 관련된 기준과 같은 사전 정보를 반영하지 않는다. 이것은 대상 지역의 특징을 직접 반영할 수 있는 장점이기도 하지만 일반화된 결론을 제시하지 못하는 단점이기도 하다. 이 문제를 해결하기 위해서는 클래스 명세서와 같은 사전 정보와 제안된 방법의 분석 결과를 결합할 수 있는 방법이 필요하며, 향후 연구를 통하여 해결하고자 한다. 제안된 방법은 적합한 유사도를 적용할 경우 면 객체 또는 선 객체의 M:N 대응 군집 쌍 탐색에서도 적용할 수 있다. 따라서 연속지적도와 수치지형도의 면 또는 선 객체 유사도를 측정하고 동일한 과정을 수행함으로써 대응 객체 군집 쌍 역시 탐색할 수 있다. 이 결과를 이용함으로써 두 지도의 기하학적 불일치를 해결할 수 있는 공액점 정보를 탐색할 수 있다. 하지만 이를 위해서는 단순한 중첩 면적 이외에 보다 정교한 유사도 측정 방법이 필요하며 이 또한 향후 연구를 통하여 해결하고자 한다.

감사의 글

본 연구는 국토해양부 첨단도시개발사업 - Hyper-Live Map 기반기술 개발과제의 연구비지원(11CHUD-C061156-01)에 의해 수행되었습니다.

참고문헌

오일석 (2008), 패턴인식, 교보문고, pp. 340-346.

허용, 김정옥, 유기윤 (2009), 지형도와 연속지적도의 가구계 폴리곤 집합간의 M:N 대응쌍 탐색, 한국공간정보시스템학회지, 한국공간정보시스템학회, 제 11권, 제 3호, pp. 47-49.

황보택근, 이기정 (2006), 시맨틱 검색을 위한 이기종 데이터간의 매칭기법, 한국콘텐츠학회지, 한국콘텐츠학회, 제 6권, 제 10호, pp. 25-33.

Bel Hadj Ali, A. (2001), *Qualité géométrique des entités géographiques surfaciques: Application à l'appariement et définition d'une typologie des écarts géométriques,* PhD dissertation, Université Marne la Vallée, Marne la Vallée.

Dhillon, I. S. (2001), Co-clustering documents and words using bipartite spectral graph partitioning, *Proceeding of 7th ACM SIGKDD Conference*, SIGKDD, San Francisco, pp. 269-274.

Duckham, M. and Worboys, M. (2005), An algebraic approach to automated geospatial information fusion, *International Journal of Geographical Information Systems*, Taylor&Francis, Vol. 19, No. 5, pp. 537-557.

Euzenat, J. and Shvaiki, P. (2007), Ontology Matching, Springer, NewYork, pp. 40-49.

Fichtinger, A., Rix, J., Schaffler, U., Michi, I., Gone, M. and Reitz, T. (2011), Data harmonisation put into practice by the HUMBOLDT project, *International journal of spatial data infrastructure research*, Vol. 6, No. 3, pp. 234-260.

Fiedler, M. (1975), A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory, *Czechoslovak Mathematical Journal*, IMAS, Vol 25, No. 10, pp. 619-633.

Huh, Y., Yu, Y. and Heo, J. (2011), Detecting conjugate-point pairs for map alignment between two polygon datasets, *Computers, Environment and Urban Systems*, Elsevier, Vol. 35, No. 3, pp. 250-262.

Hendrickson, B. (2007), Latent semantic analysis and Fiedler retrieval, *Linear Algebra and its Applications*, Elsevier, Vol. 421, No. 2-3, pp. 345-355.

Kieler, B. (2007), A geometry-driven approach for the semantic integration of geodata sets. *Proceeding of ⅩⅩⅢ International Cartographic Conference*, ICA, Moscow

Kokla, M. (2006), Guidelines on geographic ontology integration, *Proceeding of ISPRS technical commission Ⅱ symposium*, ISPRS, pp. 67-72.

Parundekar, R., Knoblock, C. A. and Ambite, J. L. (2010), Aligning ontologies of geospatial linked data, *Proceedings of the Workshop on Linked Spatio-temporal Data*, *2010*.

Pothen, A., Simon, H. D. and Liou, K. P. (1990), Partitioning sparse matrices with eigenvecotors of graphs, *SIAM Journal of Matrix Analysis and Application*, SIAM, Vol. 11, No. 3, pp. 430-452.

Sameh, A. H. and Wisniewski, J. A. (1982), A trace minimization algorithm for the generalized eigenvalue problem, *SIAM Journal of Numerical Analysis*, SIAM, Vol. 19, No. 6, pp. 1243-1259.

Trosset, M. W. and Tang, M. (2010) *On combinatorial Laplacian eigenmaps*. Technical Report 10-02, Department of Statistics, Indiana University, pp. 8-9.

Uitermark, H. T., van Oosterom, P. J. M., Mars, N. J. I. and Molenaar, M. (1999), Ontology-based geographic data set integration, *Lecture Notes in Computer Science* 1678, Springer, pp. 60-78.

Yi, S., Huang, B. and Wang, C. (2007), Pattern matching for heterogegeous geodata sources using attributed relational graph and *probabilistic relaxation, Photogrammetric Engineering & Remote Sensing*, PE&RS, Vol. 73, No. 6, pp. 663-670.

Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q. and Lin, S. (2007), Graph Embedding and Extensions: A General Framework for Dimensionality Reduction, *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, Vol. 29, No. 1, pp. 40-51.

(접수일 2012. 02. 14, 심사일 2012. 03. 26, 심사완료일 2012. 04. 14)