The Journal of GIS Association of Korea, Vol.8, No.2, pp.301-315, October 2000

시추공 정보의 클러스터링 기법을 이용한 지반분석시스템의 개발

이규병* · 김유성** · 조우석*** · 김영진****

Development of a Subsurface Exploration Analysis System Using a Clustering Technique on Bore-Hole Information

Kyu-Byung Lee* · Yoo-Sung Kim** · Woosug Cho*** · Young-Chin Kim****

요 약

지반조사 자료는 구조물을 시공하거나 설계하기 위해 필요한 기본자료이며 지층의 구성, 토질의 종류 등 지반을 구별하는 정보를 포함한다. 매년 대량으로 발생하는 지반조사 자료들은 현장의 특성을 정확하게 반영하기 때문에 지반조사 자료를 이용하여 지반을 분석한다면, 기존의 지질도나 토양도보다 뛰어난 결과를 얻을 수 있다. 따라서 비균질하고 비정형화된 지반의 특성을 고려하여 서로 구별되는 특징을 추출하고, 현장의 특성을 정확하게 반영하는 지반분석 시스템이 필요하다.

본 논문은 지반조사 정보 시스템이 관리하고 있는 시추공의 지층구성 정보와 위치 근접성을 바탕으로 시추공을 클러스터링하여 지반의 구성을 분석하는 지반분석시스템의 설계 및 구현에 대해 전반적으로 기술하였다. 개발된 지반분석시스템은 지반조사 데이터베이스의 시추공 정보를 이용하여 지반이 가지고 있는 특성 정보를 추출하고, 이를 이용하여 유사한 특성 및 위치 근접성을 갖는 시추공의 집합으로 클러스터링하여 사용자에게 정확한 지반구성 정보를 제공하는 시스템이다. 또한, 수치지도의 사용으로 지리적 위치와 지역 · 지형에 대한 지반구조의 특성조사를 가능하게 하며, 지반조사를 필요로 하는 지역에 대한 지반의 유추가 가능하여 경제적 효과를 볼 수 있다. 따라서, 지반조사 데이터로부터 다양한 종류의 정보를 얻을 수 있으며 지질도나 토양도보다 정확한 지반특성을 제공할 수 있다.

ABSTRACT : Every year, a great amount of site investigation data is collected on site to obtain sufficient information on subsurface conditions. Investigation of subsurface conditions is prerequisite to the design and construction of structures and also provides information on ground properties such as geologic formation and

* 인하대학교 전자계산학과 석사과정
** 인하대학교 전자계산학과 부교수
*** 인하대학교 토목공학과 조교수
**** 한국건설기술연구원 토목연구부 수석연구원

types of soil. This data set, which portrays real representation of ground conditions over the existing geologic and soil maps, could be further utilized for analyzing the subsurface conditions. It is therefore necessary to develope a subsurface exploration analysis system which is able to extract the valuable information from the heterogeneous, non-normalized subsurface investigation data.This paper presents the overall design scheme and implementation on a subsurface exploration analysis system. The analysis system employs one of data mining techniques, clustering technique, which extracts meaningful information out of large data set such as bore-hole data. The clustering technique employed in the developed system makes a large volume of bore-hole data into several groups in terms of ground formation and geographical vicinity. As a result of clustering, each group or cluster consists of bore-hole data with similar characteristics of subsurface and geographical vicinity. In addition, each clustered data is displayed on digital topographical map with different color so that the analysis of site investigation data could be performed in more sensible ways.

1. 서 론

인간 생활권의 주무대이며 지각의 최상부층인 지표는 암석과 암석의 부산물인 흙으로 구성되어 있으며, 거의 모든 건설공사는 근본적으로 지표 구성물질인 흙과 암석의 영향을 직·간접적으로 받게 된다. 따라서, 지반조사는 건설공사 대상 지반의 지층분포와 토질, 암석 및 암반 등 지반의 공학적 성질을 명확히 파악하여 구조물의 계획, 설계, 시공 및 유지관리 업무를 수행하는데 필수적이다.

우리나라에서는 전국적으로 매년 수많은 지반조사를 실시하고 있으며 그 결과 대량의 지반조사 자료들이 발생하고 있다. 그러나 이와 같은 지반조사 자료는 보고서의 형태로만 제출·보관되어 제한적인 재활용만이 이루어지고 있으며 유용한 많은 조사자료들이 사장되고 있다. 따라서 최근에 지반에 관련된 기존 조사자료를 수집하여 체계적으로 분류·전산화하고 통합하여 이를 간편하게 재활용하기 위한 지반조사 정보 데이터베이스 구축에 대한 연구가 진행되고 있다[1].

비균질하고 비정형화된 특성을 갖는 지반조사 데이터는 구조물을 시공하거나 설계하기 위해 이용될 뿐만 아니라, 지층의 구성이나 토질의 종류 등 지반을 구별하는 정보로 이용될 수 있다. 즉, 지반조사 데이터는 실제 지반을 천공하여 현장실험, 실내실험, 육안관찰 등을 통해서 얻어진 결과로서 지질도나 토양도([2,3])보다 현장의 특성을 정확하게 반영할 수 있다. 따라서, 지반조사 데이터에서 비균질한 지반특성을 고려하여 상호간의 관련성을 분석하고, 일관된 흐름이나 경향을 파악하여, 서로 구별되는 특징을 추출하는 방법이 필요하다. 이러한 방법의 사용은 현장의 특성을 정확하게 반영하는 지반분석을 가능하게 하고, 분석된 자료들은 다른 지반조사 방법 및 범위 설정에 이용되거나, 실시되는 지반조사량을 줄일 수 있게 하고, 현장 구성 지반의 유추를 가능케 한다. 그러나, 현재까지 연구결과는 지반조사 정보시스템을 구축하는 목적에 비중을 두고 있는 상황이므로, 이를 이용하여 지반정보를 정확하게 분석하는 시스템에 대한 연구는 미흡한 실정이다.

본 연구에서는 방대한 자료를 가지고 있는 지반조사 정보시스템의 지반조사 데이터들 간의 특성을 파악하기 위해 대용량 데이터베이스에서 데이터를 분석하는 방법 중 대표적인 기법인 클러스터링 기법을 지반조사 데이터베이스에 적용시켜 지반분석이 가능하도록 하였다. 시추공 정보에 클러스터링 기법을 적용하여 유사한

지반끼리 그룹화하고, 위치근접성을 기준으로 재분류하여 각 그룹에서 지반의 특성을 추출하고 이를 수치지도상에서 가시화 함으로써 지형, 지질 등 지반에 관한 유용한 정보의 제공이 가능한 지반분석시스템을 설계 및 구현하였다. 지반분석시스템은 지반조사 데이터의 특성을 기반으로 대용량의 데이터에 효율적으로 클러스터링이 이루어지도록 고안되었다.

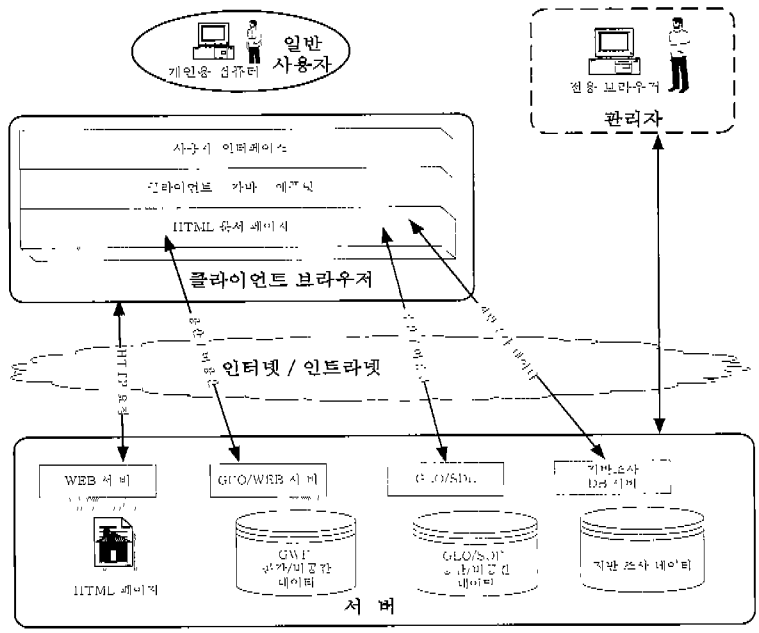
2. 관련연구

2.1 지반조사 정보시스템

지반조사 정보시스템은 지반조사 데이터를 관리하는 데이터베이스 시스템으로서 인터넷 환경에서 전용 클라이언트 및 웹 인터페이스를 통해 관리자 및 일반 사용자들이 손쉽게 이용할 수 있도록 구축되었으며 전체 시스템 구성도는 [그림 1]과 같다.

전체 시스템의 구조에서 보여주는 각 요소들은 다음과 같은 기능을 수행한다.

- 관리자용 인터페이스 : 지반조사 데이터베이스의 정보 입력 및 수정 그리고 관리를 위한 기능을 제공하며 전용 클라이언트/서버 형식을 갖고 있다.
- 일반 사용자용 인터페이스 : 일반 사용자용 인터페이스는 사용자가 웹을 통해서 지반조사 데이터베이스를 검색하는 경우에 사용된다. 웹브라우저 상에서는 두 가지의 검색이 가능하다. 수치지도에서 해당 시추공의 위치를 선택하여 검색하는 경우와 직접 검색하고자 하는 정보를 사용자가 텍스트 형태의 질의로 입력하여 검색하는 방법을 제공한다.
- 지리정보 제공 서버(GEO/Web 서버) : Geo/Web Java 클라이언트에서 의뢰된 질의에 GIS 시스템에서 질의를 처리하여 해당 데이터를 최적화된 형태로 클라이언트에게 전송하는 기능을 담당한다.



[그림 1] 지반조사 정보 시스템의 구성도

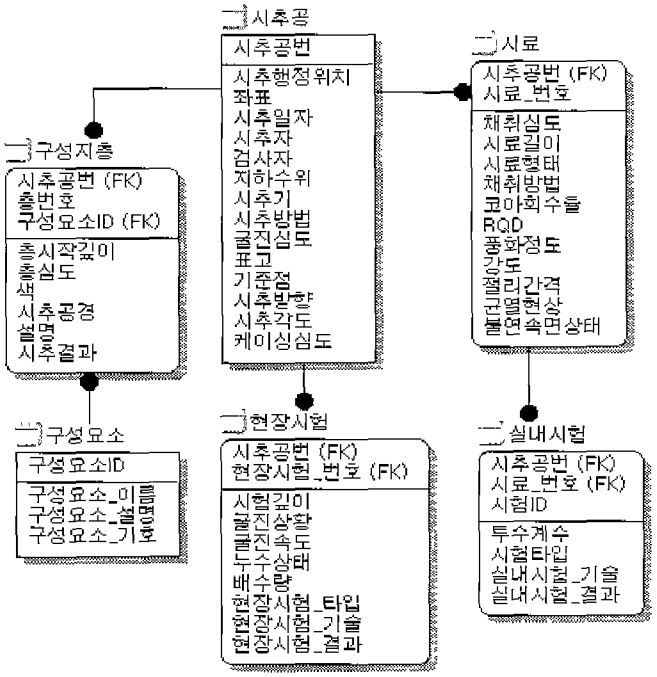
○ 원격 지리정보 등록기(GEO/Web Admin) : 데스크탑 데이터 저장구조인 GEO 포맷을 웹 데이터 저장/전송 구조인 GWF 포맷으로 변환하여 저장한다. 이러한 작업은 원격 지리정보 저장서버 호스트에 대하여도 수행될 수 있으므로 관리자가 지리정보 서비스 제공서버 호스트에 로컬 접속이 불가능한 상황에서도 웹 서비스와 원격 지리정보 등록기를 사용하여 원격으로 해당 지리정보 제공서비스를 유지 및 보수, 확장할 수 있다.

○ 데이터베이스 서버(Oracle8.x) : 지반조사 데이터를 저장하고 유지 및 관리를 위해 구축된 데이터베이스 시스템이다.

기존연구([1,5])에서 제안된 표준 지반조사 양식을 기반으로 지반조사 정보시스템을 구성하는 지반조사 데이터베이스의 E-R 다이어그램은 [그림 2]와 같다.

구축된 데이터베이스에는 시추공, 구성지층, 구성요소, 시료, 현장시험, 실내시험의 테이블로 구성되어 있다. 구성지층 테이블은 시추공을 구성하고 있는 지층정보를 구성요소별로 분리하여 한 개의 레코드로 표현한다. 구성지층 테이블의 구성항목은 층 번호, 층 시작깊이, 지층의 심도, 구성요소 정보, 시추공경, 설명, 시추결과 등으로 구성되어 있다. 지층의 구성요소를 표현하기 위해 구성요소 이름, 구성요소 설명, 구성요소 기호 등으로 구성요소 테이블이 구성되어 있다.

시추공 테이블은 시추공의 기본정보(시추공번, 시추행정위치, 좌표, 시추일자, 시추자, 검사자, 지하수위, 시추기, 시추방법, 굴진심도, 표고, 기준점, 시추 방향, 시추 각도, 케이싱 심도



[그림 2] 지반조사 데이터베이스의 E-R 다이어그램

등)에 해당하는 항목으로 구성되어 있고, 시추과정에서 실시한 시험결과를 표현하기 위해 현장시험 및 실내시험 테이블이 구성되어 있다. 현장시험 테이블에는 현장에서 시험한 시험깊이, 굴진상황, 굴진속도, 누수상태, 배수량, 표준관입시험 등의 항목이 있고, 실내시험 테이블에는 실험실에서 시험한 항목들로 구성되어 있다. 시료 테이블은 시추공에서 채취한 시료의 정보를 표현하기 위해 테이블이 구성되어 있다.

시료, 현장시험, 실내시험 테이블의 자료들은 해당 지반의 실험결과로 세부적인 성질을 제공하는 자료이다. 특히, 구성지층과 구성요소 테이블은 지반을 구성하는 여러 항목 정보의 레코드로 구성되어 있음으로, 지반을 분류하는데 필요한 정보로 사용될 수 있다.

지반정보 시스템은 인천지하철 건설본부에서 시행한 지반조사 자료를 수집, 전산화하여 구축되었다. 전국 어디서나 지반조사 자료에 쉽게 접근하여 필요로 하는 지반정보를 얻을 수 있도록 하였으며, 지반조사 결과 및 지반조사 위치정보 제공 등 다양한 서비스를 제공하고 있다[1].

2.2 클러스터링

1) 클러스터링의 정의

컴퓨터의 사용이 일반화됨에 따라 수집·보유하고 있는 데이터베이스가 양적·질적인 면에서 팽창을 거듭하면서 이를 효율적으로 사용하는 방법에 관심이 집중되고 있다. 즉 엄청난 양의 데이터를 효율적으로 활용하고 가치를 향상시키기 위해서는 데이터의 수집 이상으로 데이터의 분석이 중요한 의미를 갖는다[6,7].

대용량의 데이터를 분석하기 위한 기본적이고 대표적인 방법은 유사한 특성을 갖는 데이터들을 묶음 지워주는 클러스터링 기법이다. 클러스터링 기법에 관한 연구는 매우 활발하게 진행되어 왔으며, 실제로 수많은 클러스터링 알고리즘들이 개발되었다[8]. 이러한 알고리즘의 대부분은 클러스터 형성에 많은 시간이 소요되거나, 검색효율이나 검색시간 면에서 성능이 저조하며 클러스터의 속성이 잘 못 형성되는 문제점을 갖고 있다[9,10]. 특히 일반 클러스터링 알고리즘의 공통적인 문제점은 대용량의 데이터에 효과적으로 사용하기 힘들다는 것이다. 통계 분야에서 시작된 클러스터링 알고리즘은 소규모 데이터 집합에 사용되었으나 기술의 발전으로 대량의 데이터 집합에서 효율적으로 수행할 수 있도록 확장되었고, 그 중 몇 가지 알고리즘은 다차원 특징 벡터를 허용하고 있다[8].

일반적으로 클러스터는 유사한 성질을 가진 객체들의 집합을 말하며 집합의 유사성이란 데이터의 성질에 따라 다양해질 수 있다. 그러므로 어떤 클러스터링 알고리즘이 한 데이터 집합에 대해 최적화되어도 이 알고리즘이 다른 데이터의 집합에 최적화된다는 보장을 하지 못한다.

최근에 확장된 클러스터링 알고리즘은 유사성을 바탕으로 데이터를 그룹핑하고 데이터 집합의 중심을 찾는다. 클러스터 구성요소를 결정하기 위해 대부분의 알고리즘은 클러스터 중심과 한 점(데이터)의 차이를 평가한다[11]. 현재 대용량 데이터를 위한 기본적인 클러스터링 알고리즘은 K-Means, Linkage-based Methods, Kernel-Density Estimation Methods 등이 있다[8,12,13].

2) K-Means 클러스터링 알고리즘

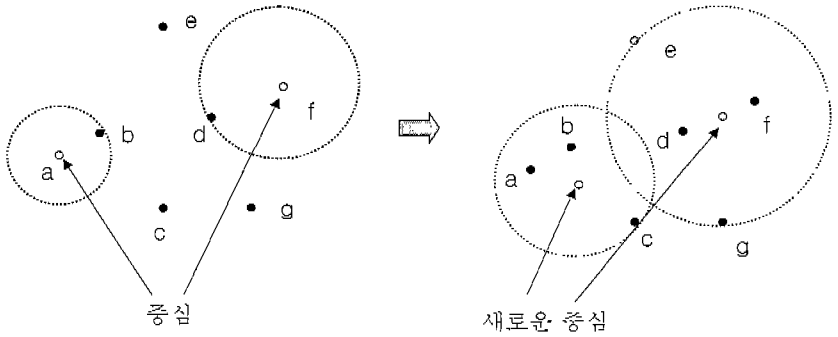
K-Means 알고리즘은 대용량 데이터 분석에 매우 우수한 효과를 보이는 알고리즘으로 데이터간의 유사성을 계산하여, 데이터들을 k개의 클러스터(cluster)로 나눈다. 즉, 하나의 데이터는 한 클러스터에 속하여 그 클러스터내의 다른 데이터와의 거리가 다른 클러스터내의 다른 데이터와의 거리보다 작다는 관계를 이용한다. 비슷한 성향을 가진 데이터들은 데이터의 공간상에서 서로 근접하게 분포할 것이므로 각 공간에서의 거리(유사성)를 근거로 k개의 그룹으로 클러

스터링하고 분할된 각 그룹에 대해 그룹내의 유사성을 중심으로 특징을 찾아낼 수 있으므로 데이터 분석에 유용하게 사용될 수 있다.

비계층적인 K-Means 알고리즘은 그래프를 분할하는 인터체인징 방법이며 원하는 클러스터의 개수가 출발점에서 가정된 분리기법이다. 각각의 클러스터링 기준이 최적화되기 위해 클러스터들 간에 점들이 배정된다. K-Means 알고리즘은 처음 시작부터 원하는 클러스터의 수와 동등한 구성요소의 수를 택함으로써 클러스터 내에 다양성을 최소화한다. 시작단계에서 요구된 클러스터의 개수 자체는 점들이 가장 멀리 떨어져 있는 점들을 선택한다. 다음 단계에서는 클러스터의 최소거리에 있는 구성요소들을 그 클러스터로 묶는다. 중심위치는 구성요소가 클러스터에 덧붙여 질 때마다 다시 계산된다. 이 과정은 모든 구성요소가 요구된 클러스터 개수에 조합될 때까지 반복된다[11].

에 속하게 되고, f와 가장 가까운 d가 f와 함께 클러스터에 속하게 된다. 중심위치는 구성요소가 클러스터에 덧붙여 질 때마다 다시 계산된다. 이 과정이 모든 구성요소가 설정된 클러스터 개수에 조합될 때까지 반복된다.

본 연구에서는 데이터간의 유사성을 기준으로 클러스터링하는 기법을 도입하였지만, K-Means 알고리즘과 동일한 알고리즘을 적용할 수는 없다. 왜냐하면, K-Means 알고리즘은 클러스터의 개수가 수행되기 전에 미리 정해지고 정해진 개수에 따라 데이터간의 유사성이 변화되기 때문이다. 따라서 지반을 구성하는 지층을 정확히 클러스터링 하는데 K-Means 알고리즘은 적당하지 않다. 본 연구에서 개발한 알고리즘은 K-Means 알고리즘의 원리를 충분히 반영하여 지반조사 정보에 효과적인 클러스터링의 결과를 얻도록 고안되었다.



[그림 3] K-Means 알고리즘의 클러스터링 과정

[그림 3]은 K-Means 알고리즘의 클러스터링 과정을 나타내고 있다.

K-Means 알고리즘은 클러스터의 수가 출발점에서 가정된다. [그림 3]에서 k가 2라면, 클러스터 수 자체는 점들 중 서로 가장 멀리 떨어져 있는 점 (a, f)을 선택한다. 점들을 선택한 다음, 선택한 점을 중심으로 클러스터의 최소거리에 있는 구성요소들을 그 클러스터로 묶는다. [그림 3]에서는 a와 가장 가까운 b가 a의 클러스터

3. 시추공 정보의 클러스터링 기법을 이용한 지반분석시스템

3.1 지반분석시스템의 개요

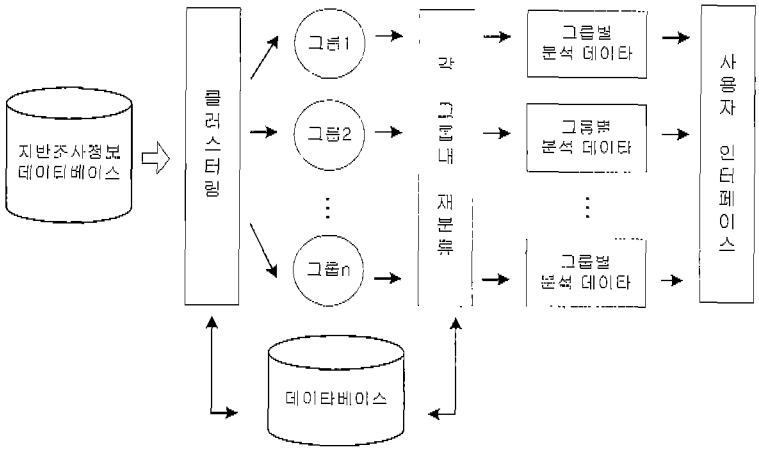
지반분석은 지반이 갖는 특성을 조사하여 지반을 구별할 수 있는 대표적인 특징 정보를 찾아내는 것이다. 시추공 정보의 클러스터링 기법

을 이용한 지반분석시스템은 지반이 가지고 있는 특성 정보 및 위치 근접성을 기준으로 지반 정보를 클러스터링하여 사용자에게 지반·지형 등 유용한 정보를 제공한다. 본 논문에서 제안된 지반분석시스템은 지반조사 정보시스템과의 연동함으로써 지반조사 정보 데이터베이스를 이용하며 지반의 특성을 추출하기 위해서 구성지층과 지층 구성비를 이용한다.

본 논문에서 개발한 지반분석시스템의 구성도는 [그림 4]와 같다.

지반조사 정보 데이터베이스에서 지반조사 데이터를 분류하기 위해 클러스터링 단계를 거치게 되고, 그 결과로서 유사한 특징을 기준으로 그룹핑된 클러스터가 생성된다. 클러스터링 단계에서 생성된 클러스터는 지반을 구성하는 구성지층이 유사한 것들로 분류되어 있고, 재분류 단계의 입력으로 사용된다. 재분류 단계에서는 지반의 비균질한 특성을 고려하기 위한 추가적인 단계로서 구성지층으로 분류된 클러스터를 다시 지반을 구성하는 지층들의 구성비와 지리적 위치근접성을 비교하여 유사한 지반들로 재분류되도록 한다.

클러스터링 단계와 재분류 단계를 거쳐 재분류된 클러스터는 유사한 지반특성을 갖는 지반들로 분류된 상태이며, 이 지반에 관한 데이터들은 구별되는 특성들로 도출된다. 이 특성들은 사용자 인터페이스 부분에서 수치지도에 표현되어 사용자에게 지반·지형 등의 유용한 정보로서 가시적인 해석을 가능하게 한다. 또한 부가적인 서비스로 시추공의 해당 시추정보를 제공하고, 그룹핑의 결과로 형성된 지반에 포함되는 시추공의 개수, 지반의 면적 등의 정보를 제공한다.



[그림 4] 지반분석시스템의 구성도

3.2 시추공 정보를 이용한 클러스터링 기법

1) 시추공 정보 클러스터링 알고리즘

제안된 시추공 정보 클러스터링 알고리즘은 유사한 대용량 데이터에 효과적인 K-Mcans 클러스터링 알고리즘을 비균질하고 비정형화된 지반의 특성을 고려하도록 변경한 알고리즘이다. 시추공 정보 클러스터링 알고리즘(<알고리즘 1>)은 지반조사 데이터에서 구성지층 정보의 시추공번과 구성요소ID를 비교하면서 지층을

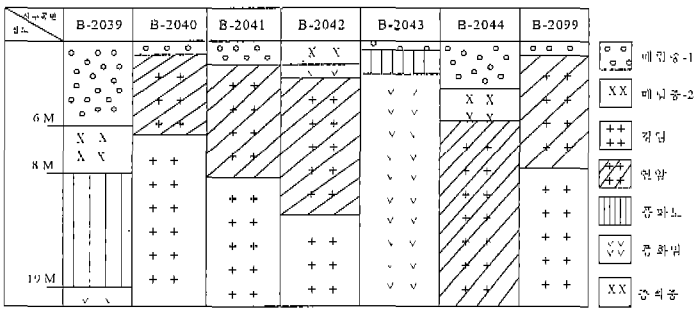
분류하게 된다. 구성요소ID는 구성요소 테이블의 데이터로서 지반을 구성하고 있는 지층의 종류를 나타내는 데이터이며, 구성요소 테이블의 한 레코드는 시추공을 구성하는 한 지층의 특성을 나타낸다.

<알고리즘 1>의 step 1에서는 지반조사 정보시스템의 데이터베이스로부터 지반조사 데이터를 입력받는다. 실제 지반조사 데이터에서 오브젝트는 시추공을 지칭하게 된다. step 2에서는 입력받을 오브젝트가 대표 오브젝트가 아닌 경우 하나의 대표 오브젝트로 선정되고 한 클러스터의 기준이 된다. step 3과 step 4는 선정할 대표 오브젝트가 없을 때까지 반복 수행된다. 알고리즘의 step 3~4는 클러스터를 생성하기 위한 단계로 대표 오브젝트와 비대표 오브젝트들간의 구성지층과 지층 구성비를 함께 비교한다. 구성지층과 지층 구성비가 유사하다면 대표 오브젝트가 속한 클러스터에 추가된다. 지층 구성비는 지반조사 정보시스템에 저장되어 있는 데이터가 아니고, 각 구성요소의 데이터에서 지층의 시작깊이와 심도의 데이터를 입력으로 받아서 계산된 값으로서 지반의 깊이를 기준으로 구성비를 산출하고 비교하게 된다. 구성비를 비교하는 이유는 지반을 이루고 있는 지층이 그 지반을 대표할 수 있을 만큼의 비율을 가지고 있는지를 판단하기 위해서이며, 비균질한 지반의 특성을 추출하기 위해서는 필수적으로 필요하다. step 4는 비대표 오브젝트들이 모두 비교될 때까지 반복하게 된다. 알고리즘의 step 5에서는 구성된 클러스터들간에 대표 구성지층과 지층 구성비를 비교하는 단계로서 유사하다면 클러스터를 통합시키게 된다. 더 이상 분류가 되지 않으면 알고리즘은 중지된다.

지반조사에 의해 얻어지는 지반의 구성지층

<알고리즘 1> 시추공 정보 클러스터링 알고리즘

step 1. 지반조사 데이터를 입력 받는다.
step 2. 더 이상 선정할 대표 오브젝트가 없을 때까지,
　　지반 데이터(오브젝트)에서 대표 오브젝트들 선정
step 3. 모든 오브젝트를 비교할 때까지,
　　대표 오브젝트와 나머지 오브젝트의 구성 지층 비교
　　If((대표 오브젝트의 구성 지층 = 나머지 오브젝트의 구성 지층) AND
　　(대표 오브젝트의 지층 구성비 = 나머지 오브젝트의 지층 구성비))
　　then 대표 오브젝트와 동일한 클러스터에 추가
　　Else step 3 수행
step 4. step 2 수행
step 5. 구성된 각 클러스터의 대표 구성지층과 지층 구성비를 이용해
　　클러스터끼리 비교. 동일하다면, 클러스터간 통합
step 6 더 이상 통합 불가능하면, 중지



[그림 5] 지반을 구성하는 지층의 예

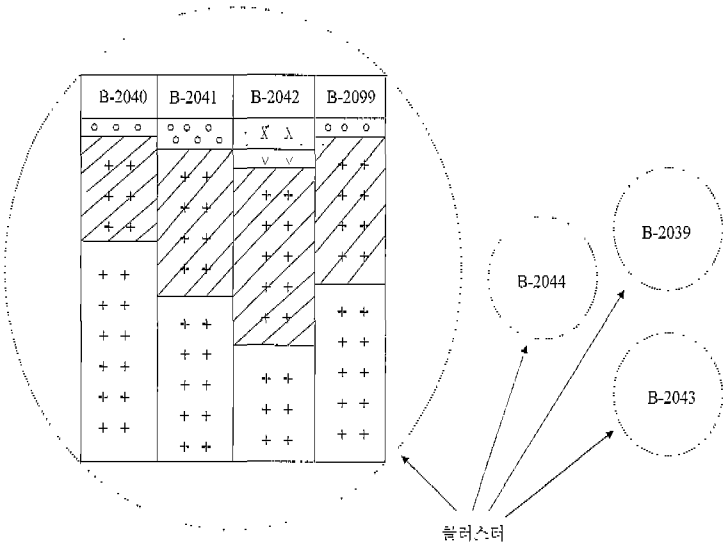
에 대한 예는 [그림 5]와 같으며, 이를 이용하여 시추공 정보 클러스터링 알고리즘을 설명하면 다음과 같다.

시추공 정보 클러스터링 알고리즘(<알고리즘 1>)을 [그림 5]에 적용시키면 먼저 구성지층간 비교과정을 거치게 된다. 대표 오브젝트를 공번 B-2041이라고 한다면, 비대표 오브젝트들은 대표 오브젝트의 구성지층을 기준으로 각각 비교되어진다. [그림 5]의 시추공 데이터들에 시추공 정보 클러스터링 알고리즘을 적용하면 [그림 6]과 같은 결과를 얻을 수 있다.

[그림 5]에서 시추공번 B-2041을 대표 오브젝트로 가정한 경우, B-2041, B-2042는 지층과 지층의 구성비가 유사함으로 동일한 클러스터로 형성되지만, 나머지 시추공들은 유사한 특징을 찾을 수 없기 때문에 각각 다른 클러스터를 생성한다. 시추공번 B-2041과 B-2042의 경우 첫 번째와 두 번째 지층은 다르지만, 전체적인 구성비를 보았을 때는 유사함으로 동일한 클러스터에 속하는 결과를 가져온다. 그러나, 비균질한 지반의 특성상 지반을 구성하는 지층이 동일한 종류의 지층으로 구성되어 있음에도 불구하고 각 지층들의 구성비가 상이한 경우 또는, 지리적으로 아주 멀리 떨어져 있는 경우에 전체 지반을 생각해 본다면 동일한 지층이라고 볼 수 없기 때문에 이러한 상황이 반드시 고려되어야 한다. 따라서, 동일 지층간에 구성비와 지리적 위치근접성을 고려하여 재분류하는 과정이 필요하게 된다.

시추공 정보 클러스터링 알고리즘에서 생성되는 클러스터는 분류된 시추공의 필요 정보만을 데이터베이스에 저장하고, 다음 재분류 단계의 입력으로 사용된다. 클러스터의 필요한 정보요소는 [그림 7]과 같다.

클러스터내의 정보요소는 재분류 단계에서 필요한 정보만을 저장하고 실제 데이터들은 지반조사 정보시스템으로부터 해당 시추공 자료



[그림 6] 시추공 정보 클러스터링 알고리즘의 적용결과 예

시추공번	X 좌표	Y 좌표	지층1의 구성비	지층2의 구성비	지층3의 구성비	지층4의 구성비	...	지층n의 구성비

[그림 7] 생성된 클러스터내의 정보요소

를 사용하게 된다. 정보요소는 클러스터에 분류되어진 시추공번, 해당 시추공의 X 좌표, Y 좌표, 해당 시추공의 지반을 구성하는 각 지층의 구성비로 구성되어 있다. 해당 시추공의 좌표를 나타내는 X, Y 좌표 항목은 동일한 지층구성을 가지고 있지만 지리적인 위치근접성을 고려하기 위해서 필수적으로 필요한 항목이며, 각 지층의 구성비는 재분류 단계에서 분류의 기준이 되는 항목이다. 지층의 구성비 항목은 지반의 깊이와 지층의 구성 수에 따라 변경될 수 있다.

2) 재분류를 위한 알고리즘

재분류를 위한 알고리즘은 생성된 클러스터에서 지층 구성비와 지리적 위치근접성을 고려하여 재분류하는 알고리즘이다. 클러스터로 분류된 지반조사 데이터는 유사한 지층들로 분류되었지만, 지층의 구성비율이 서로 상이하거나 지층과 지층의 구성비율이 유사하지만 지리적 위치가 아주 상이하다면 동일한 지반이라고 판단하기 어렵다. 따라서, 유사한 지층의 구성비를 갖는 그룹으로 재분류하기 위해 <알고리즘 2>를 이용한다.

<알고리즘 2>는 크게 두 단계로 구분된다. 첫 단계는 구성비를 이용하여 분류하는 단계이며, 두 번째 단계는 지리적 위치를 이용하여 분류하는 단계이다. 재분류를 위한 알고리즘의 step 1에서는 클러스터들을 입력으로 받게 된다. step 2는 대표 오브젝트의 각 지층의 구성비와 비대표 오브젝트간의 각 지층들의 구성비를 비교하게 된다. 이 단계에서는 지반을 구성하는 각 지층들의 구성비를 각각 비교함으로써 더 세밀한 분류를 이루게 된다. 구성비가 유사한 오브젝트들은 동일한 소 클러스터들로 분류되어진다. step 3에서는 생성된 소 클러스터들 간의 대표 오브젝트의 구성비를 비교하여 동일한 구성비를 갖게되면 클러스터를 통합하게 된다. step 4에서 더 이상 통합이 불가능하면 구성비를 기준으로 클러스터를 생성하는 분류과정을 중지하게 된다.

step 4가 종료하면 새롭게 분류된 클러스터들이 생성되며, 생성된 클러스터는 동일지층이고, 지반을 구성하는 구성비가 동일한 특성을 갖도록 분류된 결과이다. step 4를 수행한 후 생성된 클러스터들은 step 5에서 지리적 위치근접성을 기준으로 세분류하기 위한 입력으로 사용된다. step 5에서는 구성비로 분류된 클러스터의 대표 오브젝트의 좌표와 비대표 오브젝트의 좌표정

<알고리즘 2> 재분류를 위한 알고리즘

step 1. 클러스터 입력
step 2. 대표 오브젝트의 지층 구성비와 다른 오브젝트들간의 지층 구성비를 비교
  If(대표 오브젝트의 지층 구성비 = 나머지 오브젝트의 지층 구성비)
    then 그 클러스터내에서 소 클러스터들로 재분류
  Else step2 수행
step 3. 각 소 클러스터들간의 지층 구성비 비교
  구성비가 동일하다면, 소 클러스터로 통합
step 4. 더 이상 통합이 불가능하면, 중지
step 5. 소 클러스터의 대표 오브젝트 좌표와 비대표 오브젝트의 좌표비교
  가장 가까운 오브젝트를 찾아 새로운 클러스터에 추가
  D = 대표 오브젝트와 추가된 오브젝트 간의 거리
step 6. 새로운 클러스터의 모든 오브젝트와 소 클러스터의 오브젝트의 거리비교
  If(D ≥ 새 클러스터의 각 오브젝트와 비대표 오브젝트의 거리)
    then 새 클러스터에 추가
  Else step6 수행
step 7. 더 이상 추가가 불가능하면, 중지

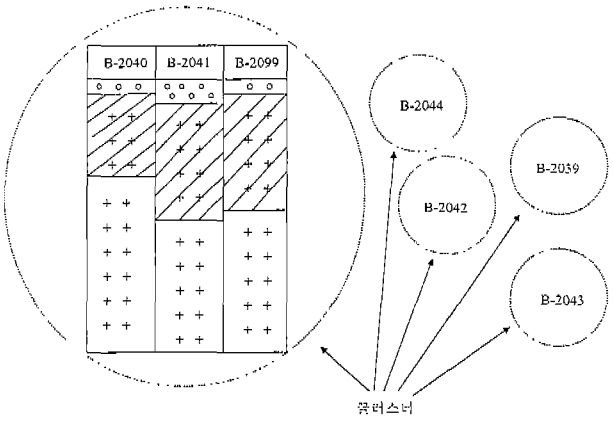
보를 이용하여 가장 가까운 오브젝트를 검색하고, 가장 가까운 오브젝트만 새로운 클러스터에 추가된다. 추가된 오브젝트는 대표 오브젝트와의 거리를 계산하게 되고, 이 거리(D)는 비대표 오브젝트와의 거리비교에 이용된다. 거리(D)를 일정한 값으로 제한한 이유는 시추조사시 일정한 간격마다 시추를 하기 때문이며 따라서, 시추공들의 간격은 대부분 동일하다는 가정에서 거리(D)를 사용하도록 한다. step 6에서는 새로운 클러스터의 모든 오브젝트와 비대표 오브젝트간의 거리를 구하여 비교해서, 거리(D)안에 속하면 새로운 클러스터에 추가된다. 새로운 클러스터의 모든 오브젝트와의 거리를 계산하기 때문에 지리적으로 위치가 근접해 있으면, 같은 클러스터에 속하게 되는 결과를 가져온다. step 7은 더 이상 클러스터에 속하는 오브젝트가 없으면 알고리즘을 중지한다.

[그림 5]의 시추공 데이타들에게 시추공 정보 클러스터링 알고리즘을 적용시켜 얻은 결과인 [그림 6]에 재분류 알고리즘을 적용하면 대표 오브젝트(B-2041)의 각 지층의 구성비와 클러스터내의 각 지층들간의 구성비를 비교하게 되고 유사한 지층으로 분류된 클러스터는 [그림 8]과 같다.

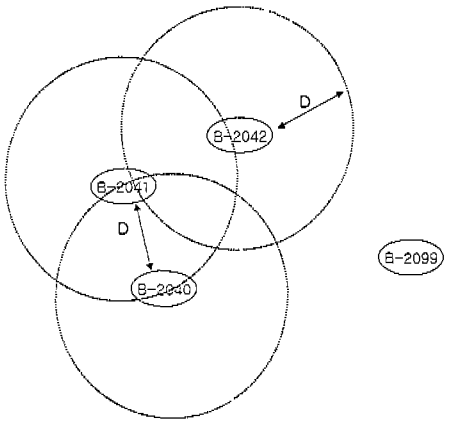
결과를 보면, 시추공번 B-2042가 클러스터에서 분리된 것을 알 수 있다. 이와 같은 결과가 나온 이유는 구성지층은 유사하나 구성지층간의 구성비가 상이하기 때문에 소 클러스터에는 추가 될 수 없기 때문이다. 따라서, 구성비가 유사한 시추공들만 새로운 소 클러스터를 생성하고, 추가된다. 두 번째 단계인 지리적 위치근접성을 기준으로 분류하는 단계를 적용시키면 [그림 9]와 같은 단계를 거치게 된다.

지리적 위치근접성 비교에서는 클러스터에 시추공들의 위치정보를 이용하여 분류하게 되며, [그림 9]에서 결과로 보여주는 클러스터내의 시추공들간의 위치정보를 비교하게 된다. 클러스터의 대표 오브젝트인 시추공번 B-2041과 지리적 위치가 가장 가까운 비대표 오브젝트인 B-2040이 새로운 클러스터에 포함되고, B-2040과의 거리(D)를 구하게 된다. 이 거리(D)는 나머지 비대표 오브젝트들과의 거리계산을 위해 사용된다. 거리(D)와 비대표 오브젝트들과의 거리계산 결과로 B-2099는 새로운 클러스터에 추가될 수 없다. B-2099는 구성지층이 유사하고 지층의 구성비도 유사하나 지리적 위치가 떨어져 있음으로 유사한 지반이라고 보기 어렵기 때문이다.

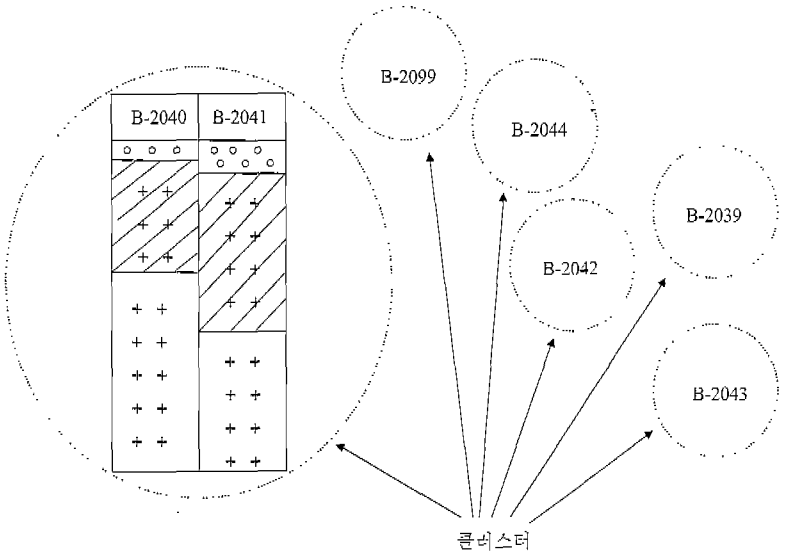
재분류 알고리즘의 두 가지 단계 중 지리적 위치비교 단계를 적용시킨 결과는 [그림 10]과



[그림 8] 구성비 비교단계 적용결과 예



[그림 9] 지리적 위치 근접성 비교단계 적용가정 예



[그림 10] 지리적 위치 인접성 비교단계 적용결과

같이 나타난다.

지리적 위치비교 단계 적용의 결과로 B-2041과 B-2040은 새로 생성된 클러스터에 속하게 되고 나머지 시추공들은 서로 특성이 다른 클러스터에 각각 속하게 된다. [그림 10]의 결과는 재분류 알고리즘의 결과로서 최종 클러스터링의 결과가 된다. 따라서 시추공 정보 클러스터링 알고리즘과 재분류 알고리즘을 모두 적용하게 되면 클러스터링이 3번에 걸쳐 이루어지며, 비균질하고 비정형화된 지반의 특성을 고려한 클러스터들이 생성된다.

3.3 지반분석 결과의 가시화

개발된 지반분석시스템에서는 시추공 정보의 클러스터링 기법에 의한 지반분석의 결과를 가시화하는 기능을 제공하여 일반 사용자가 쉽게 다차원적인 판단이 가능하도록 하였다. 데이터의 가시화는 다차원 데이터의 복잡한 관계에 대해 시각적인 해석을 제공함으로 데이터 분석 등에 유용한 방법이다[11]. 가시화된 분석정보는 클러스터링의 결과를 보여주는 수치지도와 해당 클러스터내의 데이터들과 클러스터로 그룹되어진 결과를 보여주는 대화상자 창으로 구성된다. 시추공의 위치는 지반조사 데이터의 좌표를 이용하여 수치지도에 기호로 표현되는데, 이 기호는 수치지도에서 사용하는 기호와 구별되는 특정기호를 사용한다. 해당 시추공의 기호를 선택하면 해당 시추공의 자료를 보여주게 되고, 부가적으로 해당 시추공의 지반조사 결과인 현장실험 및 실내실험 결과를 보여주도록 하였다.

클러스터링의 결과는 시추공에 해당하는 기호에 구별되는 색깔로 표현하게 되고, 동일한 색깔의 기호는 동일 클러스터내의 시추공에 해당된다. 클러스터링의 단계에 따라 시추공을 나타내는 기호의 색상속성은 바뀌어져서 수치지도에 표현된다. 클러스터링 단계의 결과로 수치지도에 표현되는 기호 외에 지반분석을 위한 클러스터링의 결과에 관한 전반적인 정보를 제공한다. 또한, 클러스터에 속한 시추공의 개수, 면적 등의 정보를 제공함으로써 구별되는 색으로 이루어진 곳의 지반특성 및 수치자료를 제공하게 된다. 결과의 표현이 수치지도상에 이루어지기 때문에 사용자에게 지반들간의 지리적 정보

까지 부가적으로 얻게 함으로써 다차원적인 분석을 가능하게 한다.

4. 지반분석시스템의 구현 및 실험

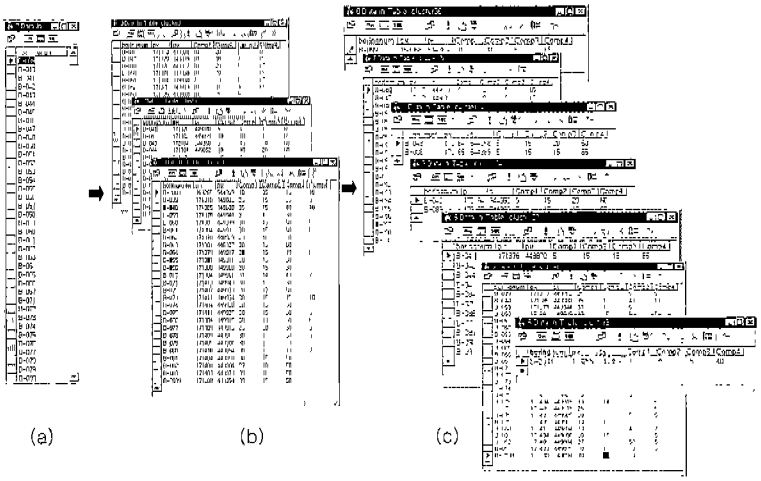
4.1 개발 환경

본 연구에서 구현된 지반분석시스템은 Windows NT4.0을 운영체제로 하는 개인용 컴퓨터 서버(Pentium-II 266MHz, 130MB RAM)에서 Microsoft Visual C++ 6.0을 이용하여 구현되었다. 지반조사 데이터를 위한 데이터베이스 관리 시스템으로는 Oracle8을 사용하였으며, GIS 응용프로그램 개발을 위해 GeoMania(v.2.5)([14])와 GDK(v2.5)([15])을 사용하였다. 그리고, 지반분석시스템의 실험을 위하여 인천시 지반조사 데이터를 바탕으로 실험데이터를 제작하였다. 인천시 지반조사 데이터를 간단함을 목적으로 수정하여 제작된 실험데이터의 지반깊이는 20m, 지층의 구성 수는 4개, 시추공의 개수는 100개로 제한하였다. 또한, 인천시 서구 가정동 일원의 수치지도를 실험에 사용하였다.

4.2 지반분석시스템의 실험

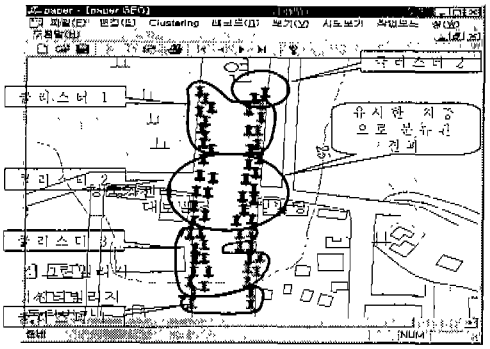
실제 실험 데이터를 이용하여 본 연구에서 개발된 <알고리즘 1>과 <알고리즘 2>를 적용한 결과 [그림 11]과 같이 클러스터를 생성하였다. 지반분석시스템의 첫 번째 단계는 시추공 정보를 클러스터링하는 단계이며, <알고리즘 1>을 적용하여 [그림 11(b)]와 같은 클러스터를 생성한다. 생성된 클러스터에는 [그림 11(a)]의 시추공 자료에서 동일한 지층으로 분류되어진 시추공들이 속해 있으며, 시추공 자료들이 지층의 종류에 따라 분류된 실제 데이터들이다.

<알고리즘 1>에 의해 생성된 클러스터에 두 번째 단계인 재분류 알고리즘 <알고리즘 2>를 수행한 결과는 [그림 11(c)]와 같다. 재분류 알고리즘은 먼저 구성비에 의한 분류를 수행하고, 지리적 위치근접성을 비교하여 최종 클러스터를 생성한다. 재분류 알고리즘은 두 단계로 나누어져 있기 때문에 실제로 클러스터링은 지층 구성비에 의한 클러스터링과 지리적 위치근접성을 고려한 클러스터링이 수행된다. 따라서, 시추공 정보를 이용한 지반분석시스템에서 지반분석 데이터는 3번의 클러스터링 과정을 거치게 된다.



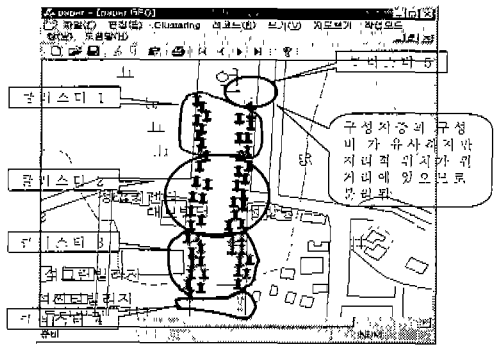
[그림 11] 클러스터링 알고리즘에 의한 클러스터 생성

본 시스템에서는 지반분석 결과를 다양한 색상의 시추공 기호로 수치지도에 나타내어 사용자의 직관적인 판단을 돕도록 하였다. [그림 12]는 시추공 정보 클러스터링 <알고리즘 1>을 적용한 결과이다. 동일한 색의 시추공 기호들은 유사한 지반이라는 것을 직관적으로 알 수 있으며, 수치지도에서 제공하는 다양한 지형정보를 통해 부가적인 분석이 가능하다.



[그림 12] 시추공 정보 클러스터링의 적용 결과

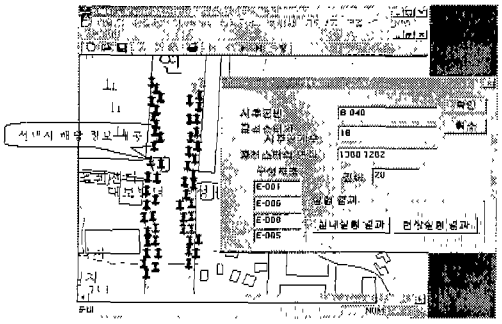
시추공 정보만을 이용하여 생성된 클러스터는 구성지층의 구성비와 지리적 근접성을 고려하지 않은 결과이며, [그림 12]에서 클러스터 2에 속하는 작은 원에 있는 지반은 큰 원 안에 속한 지반들의 특성과 유사하나 지리적 위치를 보았을 때는 같은 지반의 특성을 갖고 있다고 할 수는 없다. [그림 13]은 재분류 알고리즘인 <알고리즘2>를 적용한 결과이다.



[그림 13] 재분류 알고리즘의 적용 결과

위 그림에서 유사한 구성지층과 구성비를 갖고 있지만, 지리상 위치가 떨어져 있고, 다른 지반이 중간에 있는 경우 유사한 지반으로 간주하지 않는다. 이것은 재분류 알고리즘에서 분류기준인 구성비와 지리적 위치를 고려한 클러스터링의 결과이기 때문이다.

본 연구에서 개발된 지반분석시스템은 시추공의 개수, 시추공이 위치하는 면적, 시추공의 깊이, 구성지층, 실험결과 등 다양한 부가적인 정보를 제공한다.[그림 14]



[그림 14] 지반분석시스템에서 시추공 데이터의 서비스의 예

5. 결 론

본 논문에서는 대용량의 데이터에서 유용한 정보를 얻을 수 있는 데이터 마이닝 기법을 시추공 정보에 적용시켜 지반분석을 위한 지반분석시스템의 전반적인 연구내용을 소개하였다. 지반분석을 위해 제안된 시추공 정보의 클러스터링 알고리즘은 비균질하고 비정형화된 지반데이터와 유사한 데이터에 효과적인 K-Means 클러스터링 알고리즘을 변경한 알고리즘이다. 지반조사 데이터는 먼저 유사한 구성지층으로 클러스터링되며, 지반의 비균질한 특성에 따른 지층의 구성비와 지리적 위치근접성을 비교하여 재분류 클러스터링된다. 클러스터링 단계와 재분류 단계를 거쳐 재분류된 클러스터는 유사한 지반특성을 갖는 지반들로 분류되고, 사용자

인터페이스 부분에서 수치지도에 표현되어 사용자에게 지반·지형 등의 유용한 정보로서 가시적인 해석이 가능하도록 구현되었다. 또한 부가적인 서비스로 시추공의 해당 시추정보를 제공하고, 그룹핑의 결과로 형성된 지반에 포함되는 시추공의 개수, 지반의 면적 등의 정보를 제공한다.

본 논문에서 제안된 지반분석시스템은 지반조사 정보시스템과의 연동함으로써 지반조사 정보 데이터베이스를 이용하며, 지반조사 정보에서 지반특성을 분석하여 지질도나 토양도보다 현장 구성지반의 특성을 정확하게 반영할 수 있는 방안을 제시하였다.

향후에는 구성지층과 지층 구성비 외에 다른 인자를 고려하여 지반이 갖는 특성들의 소실을 최소화하는 효과적인 클러스터링 방법을 개발하고, 새로운 데이터 마이닝 기법을 추가하여 지층간의 연관 관계를 제시할 수 있는 연구가 필요하다.

참 고 문 헌

[1] 한국건설기술연구원, 구조물의 안전성과 경제성을 높이기 위한 지반조사 기술개발 및 D/B 구축, 과학기술부, 1999.

[2] 권영식외 5인, 지형분석, 교학연구사, 1988.

[3] 양승영, 강필종, 야외지질학, 형설출판사, 1989.

[4] 과학기술편집부, 토질조사의 계획과 적용, 과학기술, 1998.

[5] 임종석, 지반조사 상세편람, 엔지니어스, 1996.

[6] Ian H. Witten and Eibe Frank, Data Mining, Morgan Faufmann, 1999.

[7] 김정자, 이도헌, "데이터 마이닝 기술 및 연구동향," 한국정보과학회, 1998.

[8] Alexander Hinneburg and Daniel A. Keim, "Clustering Techniques for Large Data Set-From the Past to the Future," In Proceeding of KDD, 1999.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, "Knowledge Discovery in Spatial Databases," KI '99, 1999.

[10] 정영미, 정보검색론, 구미무역(주) 출판부, 1992.

[11] 박미정, "클러스터링," http://pusan.aiit.or.kr /family/student/33th/박미정 /clustering.html, 1999.

[12] Rajeev Rastogi, Kyuseok Shim, "Scalable Algorithms for Mining Large Databases," ACM SIGKDD, 1999.

[13] Raymond T. Ng and Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," In Proceeding of VLDB, 1994.

[14] GEOMania v2.5, "GEOMania v2.5 설명서," 거림 시스템, 1999.

[15] GDK v2.5, "GDK v2.5 Reference Guide for C++," 거림 시스템, 1999.