K평균 군집화를 이용한 벡터데이터 압축 기법 연구

이동헌, 전우제, 박수홍
인하대학교 지리정보공학과
ended77@yahoo.com, Woojechun@hotmail.com, shpark@inha.ac.kr

A Study on Vector Data Compression using K-means Clustering

Dong Heon Lee, Woo Je Chun, Soo Hong Park

요 약

최근 이동전화, PDA, 텔레매틱스 단말기 등과 같은 모바일 기기에서 공간데이터에 대한 사용이 증가하고 있다. 하지만 모바일 기기의 저장 공간이 늘어났음에도 불구하고 여전히 공간데이터에 대한 요구를 수용하기에는 한계가 있다. 따라서 본 연구에서는 모바일 환경에서 사용가능한 공간데이터에 대한 손실 압축 기법을 제시하고, 실험을 통한 압축률, 데이터 손실률을 분석하여 연구의 타당성과 적용 가능성을 제시하고자 한다.

세부적으로 압축률과 데이터 손실에 따르는 위치 정확도 관계에서 위치정확도를 높일 수 있는 방향을 모색하여 보았다. 그리고 다양한 군집화 기법 중 연구에 적용 가능한 기법을 선정 이용하였다. 또한 저장 공간뿐만 아닌 연산 성능 측면에서도 열악한 모바일 환경에서 만족할 만한 복원 성능을 보여야 한다. 따라서 압축된 데이터를 복원하는데 소요되는 비용을 최소화할 수 있는 방향이 연구되었다.

1. 서론

1.1 연구배경과 목적

최근 이동전화, PDA, 카 네비게이션 단말기 등 모바일 기기의 사용이 빠른 속도로 증가하고 있다. 이들 모바일 기기에서는 경로탐색, 지도 서비스 등을 위해서 공간데이터의 사용이 필수적으로 요구된다. 하지만 이런 모바일 기기는 여전히 데스크톱 환경에 비하여 제한적인 연산 수행능력과 저장 공간의 한계가 존재한다. 따라서 레스터 데이터에 비해서 상대적으로 적은 저장 공간을 차지하는 벡터데이터 조차도 여전히 큰 부담이 된다. 모바일 환경에서 사용되는 벡터형태의 데이터에는 실제로 거리 측정 또

는 경로탐색을 위한 데이터와 배경으로 사용되는 맵 데이터가 있다. 맵 데이터의 경우에는 약간의 위치 오차가 포함되더라도 그 오차정도가 눈으로 구분할 수 없을 정도의 수준이라면 받아들여 질 수 있다.[1] 몇몇 사전기반의 벡터데이터 압축 기법에 대한 선행 연구가 진행 되었지만 이들은 적용하기에는 공간 데이터의 위치 오차 수준이 너무 커지게 되거나, 오차를 줄인다면 사전이 너무 커지는 단점이 있어 실제 적용이 어려운 문제점을 가지고 있다. 따라서 본 연구에서는 맵 데이터에 대하여 K평균 군집화 기법을 이용하여 손실 압축방법을 설계 제시 하는 것을 목적으로 한다. 세부 목적으로는 사전기반의 접근방법을 이용한다. 두 번째로는 군집화 기법인 K평균 군집화를 이용하여 사전을 제작한다. 세 번째는 전체 데이터에 대한 압축률 보다는 실제 사용가능하도록 데이터의 손실률을 최소화하는 방향으로 압축한다. 마지막으로 설계된 압축 알고리듬을 선행 연구와의 비교 실험을 통하여 타당성과 적용가능성에 대하여 평가한다.

1.2 관련 연구

현재까지 데이터 압축 분야에서 벡터 데이터에 대한 압축 기법은 레스터 데이터에 비하여 연구가 다양하지 못하였다. 대표적인 사전기반의 벡터 데이터 압축에 관한 연구로는 2000년에 발표된 "Design Algorithms for Vector Map Compression"[2] 가 있는데, 이 연구에서는 벡터 데이터를 각각의 디퍼런셜 벡터로 쪼개어서 FHM(Fibonacci, Huffman, and Markov) 방법을 이용하여 미리 제작된 사전에 근사화 하는 방법을 통하여 데이터 량을 줄이는 방법을 제시 한다. 이 연구에서는 미리 제작된 사전을 사용하는 방식으로 사전을 제작하는 과정을 생략할 수 있지만 사전이 데이터의 특성을 반영할 수 없으므로 복원 시에 많은 위치오차를 갖는다. 또 다른 연구로는 데이터 특성을 찾아낼 수 있는 데이터 마이닝 기법 중 K평균 군집화 기법을 적용하여 사전을 제작하는 방법에 대하여 연구한 "Vector Map Compression: A Clustering Approach"[1] 가 있다. 이 연구의 결과는 FHM방식을 이용하는 것 보다 더 좋은 성능을 보이지만 실제 데이터에 적용 했을 때 여전히 만족할 만한 위치 정확도를 얻을 수 없다.

2. 적용 기술

2.1 사전기반 압축

사전기반 압축방법은 주어진 데이터를 대표할 수 있는 대표 값을 추출하여 사전을 제작하고 이들을 가리키는 포인터 집합을 이용하여 중복되는 값을 제거하여 전체 데이터 량을 줄이는 방법이다. 이를 여러 가지 방법으로 공간 데이터에 적용이 가능한데, 가장 간단하게 다음과 같은 적용이 가능하다. OGC에서 발표한 공간 데이터베이스의 표준인 Simple Features Specification For SQL1.1[3]방식으로 두 기하 객체를 저장하면 POLYGON((1 1, 1 2, 2 2, 2 1, 1 1)), POLYGON((1 2, 2 2, 2 3, 1 3, 1 2))와 같다. 여기에서 중복되는 점을 제거하여 POINT(1 1), POINT(2 1), POINT(1 2), POINT(2 2), POINT(1 3), POINT(2 3)와

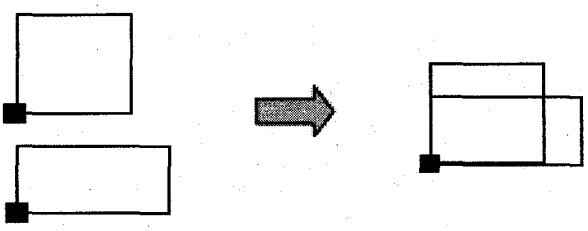
같이 6개의 좌표를 이용하여 사전을 구성할 수 있다. 이를 이용하여 폴리곤에서는 실제 좌표를 저장하지 않고 사전의 각 좌표를 가리키는 포인터만을 가지는 방식이 사전기반 압축방식이다. 연구에서는 좌표를 근사화 하였을 때 위치오차가 최소가 되는 방향으로 사전을 제작하는 것을 목표로 한다.

2.2 K평균 군집화

데이터 마이닝 분야에서 일련의 지식을 추출하는 방법의 한가지로, 미리 기준을 정하지 않고 유사도가 높은 개체를 하나로 묶어주는 방법이다.[4][6] 연구에서는 여러 가지 군집화 기법 중 K평균 군집화 기법을 사용하는데, 이는 군집 중심의 개수를 입력하여 각 데이터를 대표할 수 있는 K개의 군집으로 나누는 방법이다. 알고리듬이 비교적 간단하여 공간데이터와 같은 대용량 데이터에 적용이 가능하고, 초기 값 K 이외에 다른 사전정보를 필요치 않는 것이 장점이다.[5] 하지만 적절한 K의 크기를 정하는 것이 어렵다는 단점이 있다. 연구에서는 K 값을 변화시키면서 적절한 값을 찾고자 하였다. 또한 공간 데이터의 위치 정확도를 높이기 위하여 K 값이 커지는 것을 막기 위하여 연구에서는 개별적인 2차원 좌표를 군집화에 적용하지 않고, 길이와 각도의 두 인자로 쪼개어서 군집화 기법을 적용하였다.

3. 압축 알고리듬

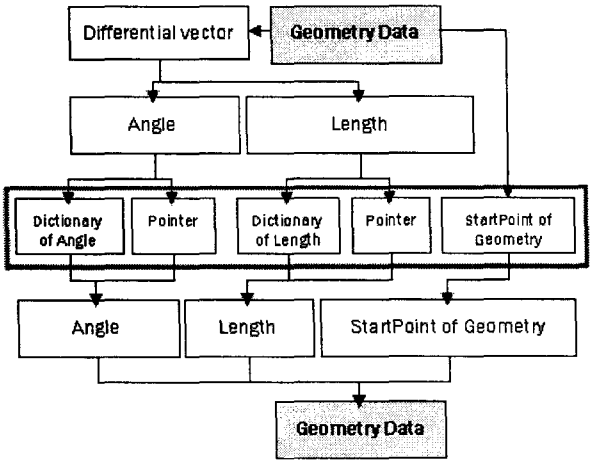
벡터 데이터를 압축하는 데에 K평균 군집화 기법을 적용하기 위해서는 기하 객체를 여러 개의 디퍼런셜 벡터로 쪼개는 과정이 필요하다. 디퍼런셜 벡터는 현재 점의 위치를 표현하기 위하여 이전 점의 좌표 혹은 해당 객체가 시작하는 점의 좌표와의 차이를 이용하여 상대적인 위치를 표현하는 벡터이다.[1] 연구에서는 개별 좌표를 해당 좌표와 이전좌표의 차이를 이용하여 디퍼런셜 벡터를 추출 하였다. 도시계획이 체계적으로 이루어진 도심지에서는 도로 중심선과 나란하도록 이전 점 좌표와의 차이 벡터가 유사한 각도를 이루며 반복해서 나타나는 경향을 보이고 있었다. 이런 점의 각도를 따로 추출하여 군집화를 수행한다면 보다 작은 K 값을 이용하더라도 좋은 결과를 얻을 수 있기 때문이다. 또한 연구에서 공간객체의 시작점을 나타내는 첫 번째 점의 좌표는 군집화를 하는 데에 포함시키지 않도록 압축되지 않은 상태로 따로 저장하게 된다. 객체의 시작점은 이전 점이 존재하지 않으므로 디퍼런셜 벡터를 만들 수 없고, 원점과의 차이를 이용하여 만들어진 벡터를 시작점들만을 따로 군집화 기법에 적용시키더라도 [그림 1]과 같이 인접 객체와 합쳐져 위치오차가 커지는 문제가 있기 때문이다.



[그림 1]인접 시작점이 같은 군집이 된 경우

다음은 이렇게 추출된 디퍼런셜 벡터를 길이와 각도로 분리하여 각각에 대하여 K평균 군집화를 적용한다. 디퍼런셜 벡터를 2

차원 공간상에서 K_1개의 군집 중심을 갖는 군집화를 적용하면 사전이 가질 수 있는 경우의 수는 K_1개가 된다. 하지만 두 개의 인자로 분리하여 각각 길이 K_2개, 각도 K_3개의 사전 두 개를 제작한다면, 전체 사전이 가질 수 있는 경우의 수는 K_2 x K_3 개가 된다. 이것은 적은 크기를 갖도록 사전을 제작하더라도 좌표를 근사화하는 단계에서 위치 오차를 줄일 수 있게 된다. 이런 과정을 거쳐 객체의 시작점, 두 개의 사전, 두 개의 사전을 가리키는 포인터의 다섯 부분으로 최종 압축된 결과를 얻을 수 있다. 위 과정의 역 과정을 통하여 위치오차가 포함된 데이터를 복원할 수 있다. [그림 2]는 일련의 압축과 복원 과정을 표현한 그림이다.

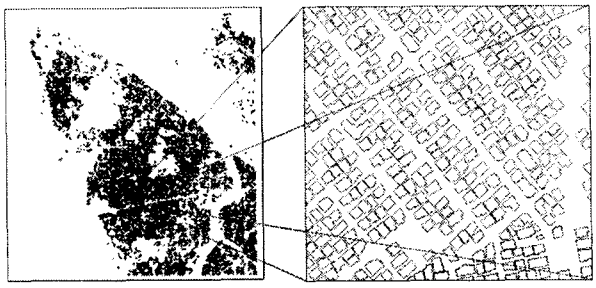


[그림 2] 압축/복원 알고리듬

4. 실험 및 분석

4.1 실험 데이터

연구지역은 서울시 양천구와 강서구 일대(8.8km x 10.1km)를 선정하였다. 실험에 사용된 데이터는 연구 지역의 1/1,000 수치지도에서 건물 코드를 갖는 지형지물을 추출한 데이터 이다. 총 72016개의 폴리곤 데이터 이며, 전체 점의 개수는 566607개 이다.



[그림 3] 실험 데이터

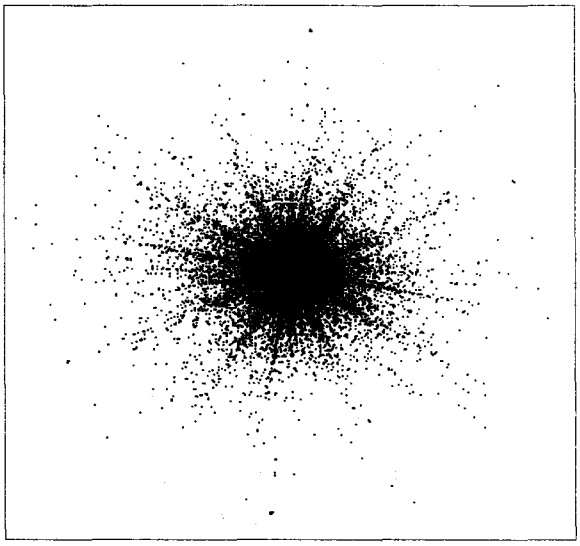
4.2 실험 내용

실험 데이터에 대하여 군집기법을 적용하기위한 디퍼런셜 벡터를 추출하였다. 연구에서 사용한 디퍼런셜 벡터는 현재 점 좌표에서 이전 점 좌표와의 차이를 이용하여 추출 하고, 각 객체의 시작점의 경우에는 원점과의 차이를 이용하였다. [그림 4]

[그림 4] 디퍼런셜 벡터

다음 과정은 이렇게 계산된 디퍼런셜 벡터에서 각 폴리곤 데이터의 시작점 벡터를 따로 저장하는 것이다. [그림 4]의 중심부분에 시작점을 제외한 나머지 디퍼런셜 벡터가

전체 점의 수에서 폴리곤수를 뺀 만큼 밀집되어있는 것을 확인할 수 있다. 데이터가 시작점을 제외한 디퍼런셜 벡터의 평균 길이보다 커지는 넓은 지역을 포함할 경우 벡터의 길이와 각도로 군집화를 수행하더라도 좋은 결과를 얻을 수 없었다. [그림 5]는 시작점을 제외한 디퍼런셜 벡터를 확대한 그림이다.

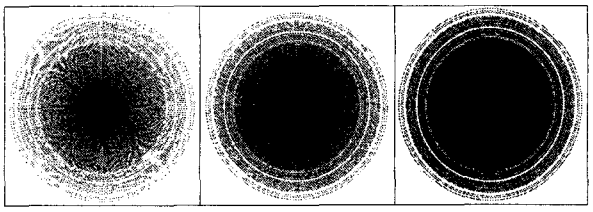


[그림 5] 시작점을 제외한 디퍼런셜 벡터

이렇게 얻어진 결과를 각각 길이와 각도로 분리하여 두 번의 K평균 군집화 기법을 적용하여 두 개의 사전과 포인터를 얻는 과정으로서 압축이 된다.

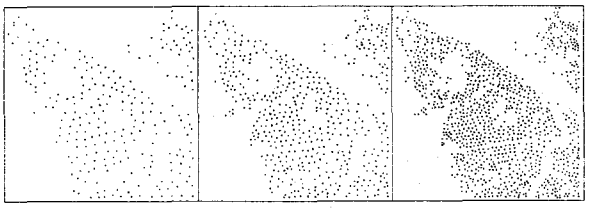
4.2 실험 결과

실험에서 K평균 군집화 과정은 통계 패키지인 SPSS v10 을 이용하여 수행 하였다. 초기 값 K를 거리와 각도 모두 256, 512, 1024 개로 늘려가면서 군집화 하였다. 이렇게 해서 생성된 거리와 각도의 사전으로 조합 가능한 경우의 수를 표현해 보았다.[그림 6]



[그림 6] 조합 가능한 벡터 사전

[그림 6]에서 좌로부터 각도, 길이의 사전 수가 각각 256, 512, 1024 개로서 가질 수 있는 모든 경우의 수는 65536, 262144, 1048576 개가 된다. 이것은 [그림 5]에서 나타나는 디퍼런셜 벡터가 [그림 6]의 사전에 가장 가까운 값으로 근사화 될 때 사전이 클수록 적은 오차를 포함할 수 있는 것이다. [그림 7]은 선행연구 방법을 실험 데이터에 적용 하였을 때 가질 수 있는 사전의 분포이다.



[그림 7] 선행연구의 벡터 사전

[그림 7]은 각각 좌로부터 256, 512, 1024개의 벡터를 갖는 사전을 표현한 것이다. 이는 동일한 데이터 량을 갖는 사전을 두 가지 다른 방식으로 제작 하였을 때 사전이 표현할 수 있는 벡터수의 차이이다. 각 디퍼런셜 벡터를 사전에 근사화 하는 방법으로 압축된 데이터를 복원 과정을 거쳐 원본과의 비교를 통해 데이터 손실에 따르는 위치 오차 정도, 최종 데이터 크기, 압축률을 계산해 보았다.[표 1] 각각 사전 크기를 65536, 262144, 1048576 개로 늘려가며 실험

하였다. 압축된 데이터의 크기는 폴리곤의 시작점을 저장하는 부분, 디퍼런셜 벡터의 거리와 각도를 대표하는 두 개의 사전, 사전을 가리키는 두 개의 포인터의 다섯 부분을 합한 것이다. 또 압축률은

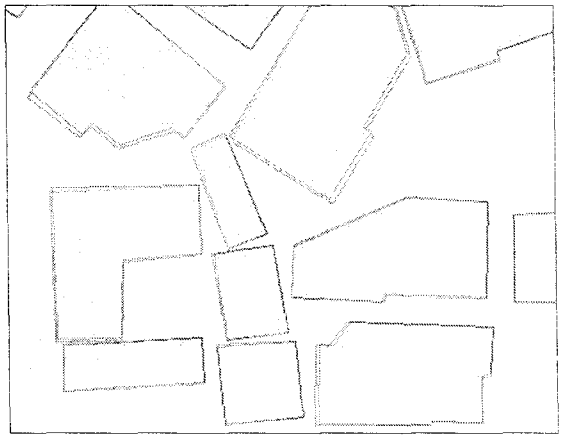
$$\text{압축률}(\%) = \frac{\text{원본데이터크기} - \text{압축데이터크기}}{\text{원본데이터크기}} \times 100$$

의 식으로 계산하였다. 원본 데이터의 크기는 9065712byte 이다.

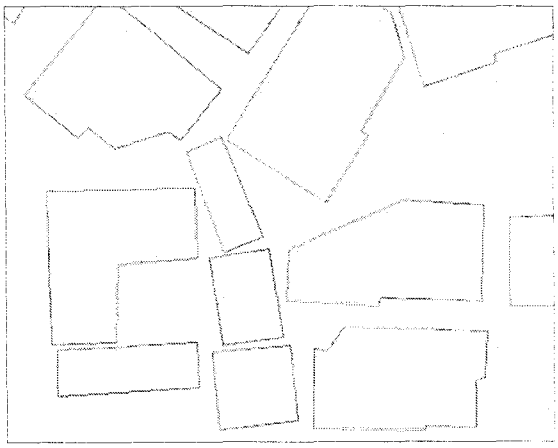
[표 1] 결과 분석

	실험 1	실험 2	실험 3
RMSE(m)	0.048840	0.010844	0.002756
Size(byte)	2001502	2111241.75	2225077.5
압축률(%)	77.922286	76.711793	75.45612

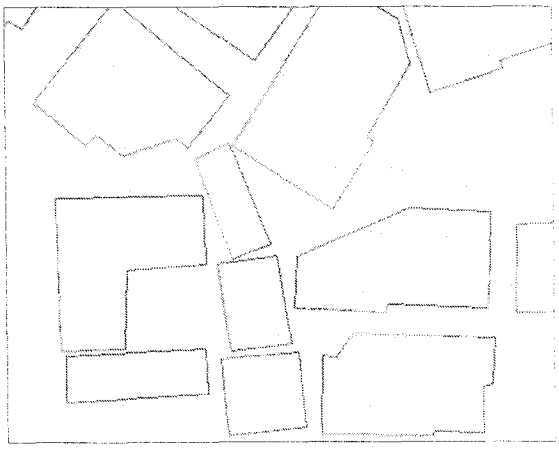
[표 1]에서 세 번의 실험에서 각각 위치 오차 수준이 상당히 작은 수준으로 나타나는 것을 볼 수 있다. [그림 8], [그림 9], [그림 10]은 각각 길이와 각도의 사전 크기를 늘려가면서 실험한 결과이다. [그림 10]의 경우 데이터 크기를 약 25%정도로 줄이면서 정확도를 유지하는 결과를 보였다.



[그림 8] 사전 크기 256 x 256 경우



[그림 9] 사전 크기 512 x 512 경우



[그림 10] 사전 크기 1024 x 1024 경우

연구에서 제시한 방법의 결과와 유사한 결과를 보이도록 선행연구의 사전크기를 조절하였다. 또한 인접한 두 객체가 합쳐지지 않도록 시작점을 사전에 포함하도록 제작하여 실험하였다. 실험 결과는 [표 2]와 같다.

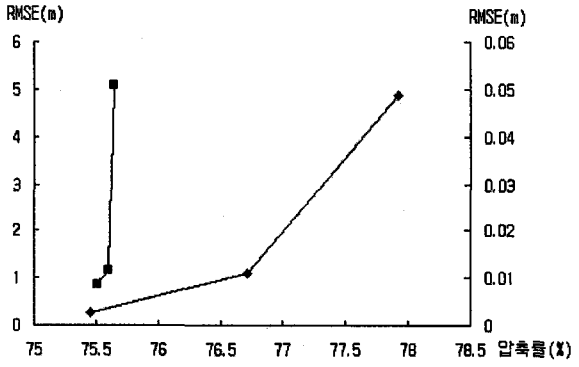
[표 2] 선행 연구 결과 분석

	실험 1	실험 2	실험 3
RMSE(m)	5.086660	1.144022	0.860976
Size(byte)	2207357.875	2211453.875	2219645.875
압축률(%)	75.65155773	75.60639611	75.51603366

3번의 실험에서 사전의 크기를 72272,

82528, 73040 개로 늘려가면서 실험을 하였다. 사전은 시작점 72016개를 포함하고 각각 256, 512, 1024 개에 초기값을 이용하여 K평균 군집화를 통한 결과이다.

선행 연구의 결과와의 비교에서 비슷한 수준의 압축률을 보이는 경우 더 향상된 위치정확도를 보이고 있다.[그림 11]



[그림 11] 압축률과 위치오차 관계

5. 결론

본 연구에서는 K평균 군집화 기법을 적용한 사전기반의 벡터 압축 방법을 설계하고 실험을 통하여 적용 가능성을 타진해 보았다. 하나의 객체 내에서 상대적 위치를 나타내는 디퍼런셜 벡터의 다양한 경우를 포괄하기 위하여 길이와 각도 두 인자로 분리하여 군집화를 수행하고, 좌표계 내에서 절대적인 위치를 정의하기 위하여 각각의 공간 객체에서 하나의 점, 연구에서는 시작점을 변형을 가하지 않고 그대로를 저장함으로서 압축률을 떨어트리지 않고 위치정확도를 향상시킬 수 있었다.

연구에서 제시된 방법을 적용한다면, 벡터 데이터에 대하여 손실을 최소화 하면서 데이터의 크기는 약 25% 수준으로 낮출 수 있을 것으로 기대된다.

6. 참고 문헌

[1] Shashi Shekhar, Yan Huang, Judy Djugash, Changqing Zhou, "Vector Map Compression: A Clustering Approach", Proceedings of the tenth ACM international symposium on Advances in geographic information systems, 2002, 74 - 80

[2] D. Salomon. "Data Compression: the Complete Reference." Springer-Verlag, 2nd edition, 2000.

[3] David Beddoe, Paul Cotton · Robert Uleman, Sandra Johnson, Dr. John R., Herring, "OpenGIS Simple Features Specification for SQL Revision 1.1", OpenGIS Consortium., 1999

[4] Macqueen, J. "Some methods for classification and analysis of multivariate observations.", In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 281 - 297.

[5] A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Volume 31, Issue 3, 1999, 264 - 323

[6] Jiawei Jan, Micheline Kanber, Datamining comcepts and Techniques", Morgan Kaufmann, 2000