K-means 알고리즘을 사용한 분산 바이오 데이터 통합화

류병걸○ 신동규
신동일 정종일
baramryu@gce.sejong.ac.kr, shindk@sejong.ac.kr,
dshin@sejong.ac.kr, jijeong@gce.sejong.ac.kr

Integration of Distributed Biological Data using Modified K-means Algorithm

Byunggul Ryu○ Dongkyoo Shin
Dongil Shin Jongil Jeong

요 약

Bioinformatics의 목표는 생물학적인 질의를 해결하는 것과 생물학자들이 수집된 데이터를 분석하고 검색을 하여 생물학자들이 정확한 일을 수행하는 것이다. 인터넷은 여러 조사 그룹의 데이터베이스에 동시에 접근가능한 수단을 제공했으나 이러한 분산 환경에서 많은 양의 데이터는 전송 시의 시간 지연 문제와 최종 검색시의 느린 검색 속도 문제를 나타낸다. 데이터 클러스터링은 데이터의 검색시 이러한 문제점을 해결하기 위하여 이용될 수 있는 방법이지만 단순 적용시에는 데이터의 양에 비례하는 실행 시간이 또다른 문제를 발생시킨다. 본 논문에서는 바이오데이터의 효율적인 클러스터링을 위한 개선된 분산 클러스터링 시나리오와 이를 위해 수정된 K-means 알고리즘을 제시한다. 최종 실험 결과는 20% 이상 향상된 실행 속도를 보여준다.

1. 서 론

인터넷의 확산으로 분산환경하에서 많은 양의 데이터를 전송하는 것은 쉬워졌다. 특히 Bioinformatics 영역에서 인터넷의 발전은 서로 다른 연구 그룹간에 바이오 데이터를 공유하는 것이 가능해져서 생물학적 질의를 해결할 수 있는 많은 가능성을 나타냈다[1]. Bioinfomatics의 목표는 생물학적인 질의를 해결하는 것과 생물학자들이 정확하게 연구할 수 있게 하는 것이다. 그럼에도 불구하고, 그 안에는 해결해야 할 어려움들이 많이 있다. Bioinfomatics의 많은 데이터는 분산환경하에서 데이터 전송과 데이터를 검색할 때 시간을 지연시키는 문제를 발생한다.

본 논문은 5장으로 구성되어졌다. 2장에서는 데이터 클러스터링에 관한 K-means 알고리즘과 microarray 데이터를 XML로 표현하는 MAGE-ML에 대하여 소개한다. 3장에서는 bioinfomatics 데이터의 통합과 교환을 위한 시나리오와 수정된 K-means 알고리즘을 설명한다. 4장에서는 제시된 시나리오를 측정하고, 마지막 5장에서는 결론을 서술한다.

2. 배 경

현재 Bioinformatics를 연구하며 발생하는 문제들을 해결할 수 있는 여러 연구결과가 있다. 대표적인 연구는 분산 데이터베이스의 응용과 클러스터링 알고리즘에 관한 것이다.[2][3] 이 연구들은 e-business분야뿐만 아니라 bioinformatics 분야에서도 매우 효율적이다. 다음 세부 파트에서는 bioinformatics 데이터 클러스터링을 위해 주로 사용한 K-means 알고리즘과 microarray에 관한 자료를 설명하고 microarray를 전송하기 위해 구성된 언어인 MAGE-ML (MicroArray Gene Expression Markup Language)[4] 을 소개한다.

2.1 K-means

클러스터링은 분산환경에서 각 각 존재하는 데이터를 분석하고 검색할 때 사용된다. K-means는 매우 유용한 클러스터링 알고리즘이다. K-means 알고리즘은 다음 단계를 따라서 수행된다[5].

- 단계 1
데이터 묶음[X1...XN]에서 랜덤하게 선택하여 k개의 백터[Y1……YK]을 초기화 한다.

- 단계2

본연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어짐. (0412-M101-0416-0002)

XN이 Yi에 가장 가깝다면 Xn을 Yi에 속한다고 표시하고 모든 데이터 묶음 [X1……XN]이 K 갯수의 클러스터에 나뉘어 진다.
Xi = {Xn | d(Xn, Yi) ≤ d(Xn, Yj), j = 1,…K}

- 단계 3
2 단계에서 생성된 클러스터들의 중심점을 새로 연산하여 갱신한다.
Yi = c (Xi), i = 1,…K

- 단계 4
가장 가까운 클러스터의 중심과 데이터들의 거리의 합을 계산하여 총 왜곡을 구한다.

$$D = \sum_{m=1}^{N} d(x_n, y_{i(n)})$$

where, i(n) = k, if $x_n \in X_k$

- 단계 5
왜곡률이 더 이상 변화하지 않거나, 정해진 반복횟수가 될 때까지 2~4단계를 반복한다.

2.2 MAGE-ML

MAGE(MicroArray Gene Expression)는 3부분으로 구성된다 : MAGE-OM(Object Model), MAGE-ML (Markup Language), and MAGE-STK (MAGE-Software Tool Kit). MAGE-OM은 microarray 데이터의 표준 객체 모델이고, MIAME에 132개의 클래스와 17개의 패키지로 구성되어졌다. MAGE-OM을 기반으로 만들어진 MAGE-ML은 microarray 데이터를 다른 곳으로 전송하기 위해 XML을 기반으로 한 표준 데이터 양식이다. MAGE-STK는 MAGE-OM을 실행하기 위해 여러 프로그램 언어로 나타낸 것이다[6].

MAGE-ML은 <MAGE-ML>로 쓰여진 하나의 루트 엘리먼트를 가지고, 루트 엘리먼트는 13개의 자식 엘리먼트를 갖는다[1],[4],[6]. 각 엘리먼트는 또 여러 자식 엘리먼트를 가질 수 있다. 루트 엘리먼트는 모든 엘리먼트를 가지거나 선택적으로 가질 수 있다.

- Experiment : microarray 실험에 관련된 정보
- ArrayDesign : array 디자인과 관련된 정보
- BioMaterial : microarray 실험 소재와 관련된 정보
- BioAssayData : microarray 요소와 관련된 정보
- BioSequence : 순서와 관련된 정보
- QuantitionType : microarray 실험후 측정 관련된 정보.
- Array : 각 array에 관련된 정보
- Protocol : microarray실험을 위해 사용된 소프트웨어와 하드웨어에 관련된 정보
- AuditAndSecurity : 연구실 또는 연구자에 관한 정보
- Description : 실험에 관한 의견과 첨부된 파일과 관련된 정보.
- HigherLevelAnalysis : 실험 데이터의 분석에 관련된 정보

MAGE-OM/ML에서 제안한 디자인은 다른 연구자들과 microarray 데이터를 공유하기 위한 것이다. 각 그룹들이 동일한 MAGE 데이터베이스 시스템을 사용한다면 각 연구자들은 다양한 마이닝 기술을 이용하여 정확한 microarray 데이터를 수집할 수 있다. 이런 경우처럼 MAGE에서 제안된 디자인을 적용하면 다른 특정한 단체에 속해 있는 그룹으로부터 도출된 데이터를 좀 더 쉽게 가져오고 통합할 수 있다.

3. 분산환경하에서 대용량의 bioinformatics 데이터의 통합과 클러스터링

이 파트에서 우리는 분산환경하에서 bioinformatics 데이터의 통합을 위한 효율적인 시나리오를 제시한다.

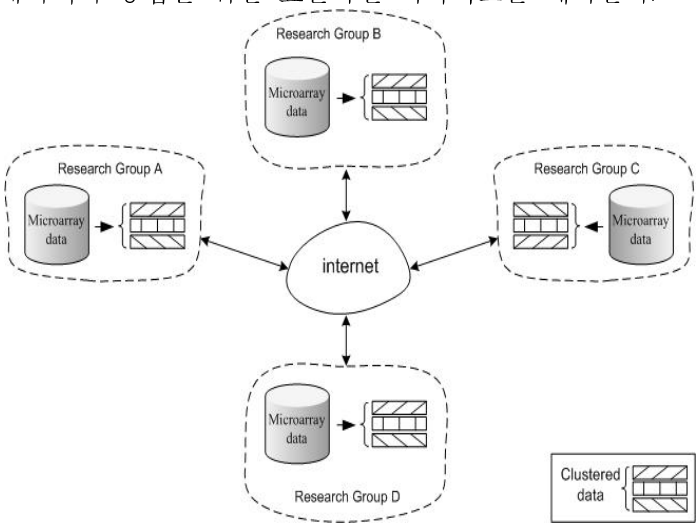

그림 1. Distrubuted repositories for microarray data

그림 1은 몇몇 연구 그룹이 인터넷을 통해 연결된 실험환경을 보여준다. 기본 시나리오는 다른 연구 그룹으로부터 bioinformatics 데이터를 얻고, 그것을 하나의 연구그룹이 통합하는 것이다. 이 시나리오는 우리가 제시한 데이터 클러스터링과 XML 압축이 핵심 기술이다. 핵심 기술을 적용하기 위해 다음고k 같은 시나리오를 구성했다.

- 시나리오 1 : 각 그룹(A ,B, C)은 그룹 D에게 bioinformatics 데이터를 전송한다. 그룹 D는 각 그룹으로부터 가져온 데이터를 통합하여 K-means를 써서 클러스터링 한다. 클러스터링은 그룹 D에서 한번만 실행된다.
- 시나리오 2 : 각 그룹(A, B, C)은 그룹 D에게 bioinfomatics 데이터를 요청 받으면 각 그룹별로 클러스터링 하고, 데이터와 클러스터링의 중심값을 전송한다. 그룹 D는 각 그룹에서 전송된 중심값의 평균을 구하고, 그 평균 중심값을 가지고 전송된 데이터를 통합하여 클러스터링한다. 이 클러스터링

부분에서는 2.1장에서 소개된 K-means 알고리즘의 단계 1을 수정하여, 초기 중심값으로 평균 중심값을 사용했다. 그리고 2.1장에서 보여준 나머지 단계를 수행했다.

이 시나리오들을 실행하면서 우리는 시나리오2가 클러스터링, 전송, 분산환경하에서 bioinformatics의 방대한 양의 데이터의 통합과 수행능력이 개선된 것을 알 수 있었다.

4. 평가

이번 장에서는 3장에서 보여준 두 시나리오의 평가를 위해서 실험을 했다. 여기서는 physical bioassay에서 가져온 속성 값들이 있는 bioassay 데이터를 이용하여 실험을 했다. 가져온 bioassay는 MicroArray 실험 단계중 하나인 BioArray로 나타낸다. 측정된 bioassay 데이터는 40개의 속성을 가지고 있고, 7개의 속성 (CH1I_MEAN, CH2I_MEAN, CH1B_MEDIAN, CH2B_MEDIAN, CH1D_MEAN, CH2D_MEAN, and CH1I_MEDIAN [6])을 선택해야 하는 두 개의 클러스터 그룹을 만들어 K-means 클러스터링을 수행했다.

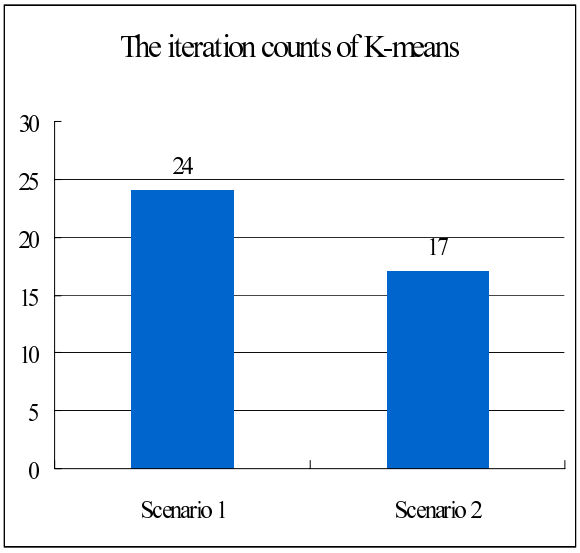

그림 2 : comparison of iteration counts in research group D

그림 2는 두 시나리오의 총 클러스터링 반복 횟수를 보여준다. 시나리오 1처럼 연구그룹 D가 다른 그룹(A, B, C)에게 데이터를 요청하고 각 연구그룹들은 연구그룹 D에게 데이터를 보내서 연구그룹 D만 클러스터링을 할 경우에 24번 반복하게 된다. 시나리오 2처럼 연구그룹 D가 데이터를 요청하면 다른 그룹(A, B, C)이 각 각 클러스터링후 연구그룹 D에게 평균 중심값을 보내줘서 연구그룹 D가 클러스터링을 할 경우 17번 반복하게 된다.

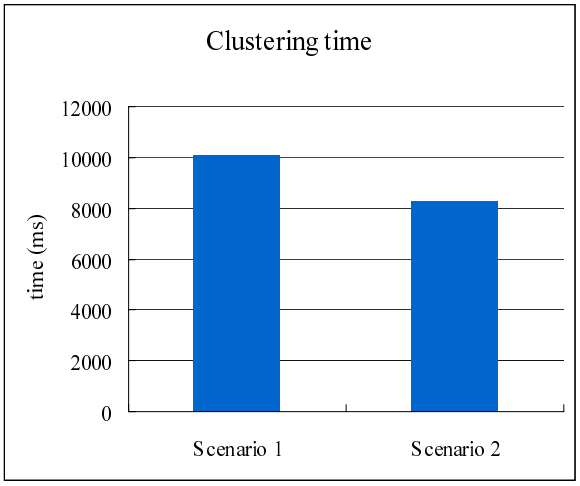

그림 3 : A comparison of clustering time in research group D

그 결과 수정된 K-means알고리즘을 사용한 시나리오 2에서 성능이 개선된 것을 알 수 있다. 또한, 여기서 두 시나리오의 K-means 실행시간을 밀리세컨드 단위로 측정했다. 그림 3에서 보여지는 것처럼 실행 시간은 약 20% 단축되었다.

Data	Centroid						
	Att1	Att2	Att3	Att4	Att5	Att6	Att7
Cluster 0							
Research Group A	1597.866	1686.334	125.5074	154.9082	1472.358	1531.426	1522.416
Research Group B	642.7346	641.1411	68.6306	142.7207	574.1039	498.4204	605.5098
Research Group C	1258.544	1113.168	117.6433	98.3927	1140.911	1014.776	1189.647
Averaged centroid	1166.385	1146.881	103.9271	132.0072	1062.458	1014.874	1105.857
Final centroid	1221.97	1194.27	103.641	133.294	1118.32	1060.97	1162.92
Cluster 1							
Research Group A	1178.88	8743.522	129.8076	160.5319	11049.07	8582.99	11132.37
Research Group B	3677.535	2489.578	75.0644	171.0307	6302.47	2318.547	3660.108
Research Group C	10799.72	8389.727	129.0404	106.001	10670.68	8283.726	10599.97
Averaged centroid	8552.042	6540.942	111.3041	145.8545	8440.738	6395.088	8464.15
Final centroid	10181	7352.93	121.441	140.915	10059.6	7212.02	10080.5

Table 1 : Centroids of attributes in Scenario 2

테이블 1은 각 연구그룹(A, B, C)에서 클러스터링을 하

여 측정된 중심값과 평균 중심값 그리고 연구그룹 D에서의 최종 중심값을 보여준다. 클러스터 0과 1의 각 속성들의 평균 중심값이 최종 중심값과 유사하다는 것을 알 수 있다. 여기서 각 클러스터의 평균 중심값을 초기 중심값으로 사용하기에 매우 적합하다는 것을 알 수 있다.

5. 결론

본 논문은 분산 환경하에서 바이오 데이터 클러스터링을 위한 효율적인 분산 클러스터링 시나리오와 수정된 K-means 알고리즘을 제시했다. 제시된 방법에 의하면 클러스터링 반복횟수는 30% 이상 감소되었고, 클러스터링 실행 시간은 20% 이상 개선 되었다. 향후에는 좀 더 개선된 분선 클러스터링 알고리즘에 대한 연구를 진행할 예정이다.

6. 참고자료

1. Isabelle Crignon, Stefan Grzybek, Frank Staedtler, Anthony Masiello, Marlene Dressman, Fred Taheri, Rainer Stock, Elodie Lenges, Rosario Pitarelli, Fabrizio Genesio and Mischa Reinhardt: An Architecture for Standardization and Management of Gene Expression Data in a Global Organization, ECCB (European Conference on Computational Biology) 2003, Paris, France, September 27-30, 2003.
2. B.Zhang, G. Formaml: Distributed Data Clustering Can Be Efficient and Exact, Software Technology Laboratory HPLaboratories, Palo Alto HPL-2000-158 December 4th, 2000
3. O. Albert Y. Zomaya, Tarek E1-Ghazawi.:Parallel and Distributed Computing for Data Mining, IEEE Concurrency, vol. 7, No. 4, Pages:11-13, 1999
4. MAGE-ML, http://www.mged.org/Workgroups/MAGE/mage-ml.html
5. Yu-Fang Zhang; Jia-Li Mao; Zhong-Yang Xiong, An efficient clustering algorithm, 2003 International Conference on Machine Learning and Cybernetics, Volume 1, Pages:261 - 265, Nov. 2-5 2003
6. William Martin, Robert M.Horton : MageBuilder: A Schema Translation Tool for Generating MAGE-ML from Tabular Microarray Data, 2003 IEEE Proceedings of Computational Systems Bioinformatics (CSB '03), Pages:431 - 432, 2003.