다차원 데이터큐브의 근사 질의응답 기법

이선영*, 김영주*, 배우식*, 이종연*
*충북대학교 컴퓨터교육과
e-mail:elesun97@chungbuk.ac.kr

The Approximate Query Answering Method in Multi-dimensional Data Cube

Sun Young Lee*, Yeong Ju Kim*, Woo Sik Bae*, Jong Yun Lee*
*Department of Computer Eduction, Chungbuk National University

요 약

DSS 응용들의 대용량 집계 데이터 집중 시스템에서는 효율적이고 즉각적인 의사결정 지원을 위한 근사 질의응답의 연구가 필요하다. 따라서 본 연구에서는 FCM 클러스터링 기법과 ANFIS을 이용한 기법을 제안한다. 제안된 기법은 다차원 데이터 큐브의 데이터 특성을 가지며 질의에 대한 근사적인 응답을 제공할 수 있는 모델을 생성한다. 제안된 기법을 통해 학습된 모델은 기존의 기법보다 근사 질의응답의 정확성이 향상되었음을 비교 실험을 통하여 확인한다. 따라서 제안된 기법은 기존의 기법보다 저장 공간과 시간을 줄일 수 있으며 또한 근사 응답의 정확도를 향상시킬 수 있다.

1. 서 론

데이터웨어하우스는 이기종, 운용 또는 legacy 시스템으로부터 통합되고 추출된 데이터의 전역 저장소로 이력 정보를 유지하고 시간에 따라 성장한다. 또한 전문가, 매니저, 분석가들이 더 빠르고 더 나은 결정을 만드는 것을 돕기 위해 구축된 집계 데이터의 커다란 집합이며, 의사결정 지원시스템(Decision Support System, DSS)에서 가치 있고 유용한 지식을 추출하기 위해 사용된다[1]. 데이터웨어하우스의 실제 물리적 구조인 다차원 데이터 큐브는 데이터에 대한 다차원적 시각을 제공하며, 요약 데이터의 사전계산과 신속한 접근을 가능하게 한다.

1.1 연구동기

DSS에서 지식을 추출하기 위해 데이터웨어하우스에 질의를 하면 정확한 응답을 제공하기 위해서 많은 응답 시간을 소비한다. 하지만 DSS 응용들은 데이터들에 대한 꼼꼼한 분석 보다는 경향분석에 더 많은 관심을 가지며, 정력적이기 보다는 정성적인 분석질의로 소수점 이하의 정답을 필요로 하지 않는다[2, 3]. 따라서 대용량 집계 데이터 집중 시스템에서는 효율적이고 즉각적인 의사결정 지원을 위한 근사 질의응답(Approximation Query Answering, AQA)의 연구가 필요하다. 또한 기존 AQA 방법들은 주로 데이터 큐브를 압축하고, 이 압축된 큐브를 기반 AQA를 제공하는 방식을 채택하고 있다[2]. 그러나 이 기법들은 저차원의 데이터 큐브에서는 수행능력이 뛰어나지만 고차원의 경우 많은 계산을 요구한다. 또한 고정된 계층화 차원을 갖는다면 집계 질의는 미리 정해진 차원에 대해서만 가능하다. 따라서 기존 AQA 기법들보다 정확성이 우수하며 사전계산된 데이터의 저장량이 적은 기법을 제안하는 것이 필요하다.

1.2 기여도

본 논문에서는 Fuzzy C-Means(FCM) 클러스터링 기법과 Adaptive Neuro Fuzzy Inference System (ANFIS)을 이용해 다차원 데이터큐브의 지능형 모델을 생성한다. 학습된 모델은 다차원 데이터 큐브의 데이터 특성을 가지고 있으며 질의에 대한 근사적인 응답을 제공할 수 있다. 또한 적은 파라미터로 데이터를 표현함으로써 다차원 데이터 큐브의 압축된 표현을 저장할 공간을 줄일 수 있다.

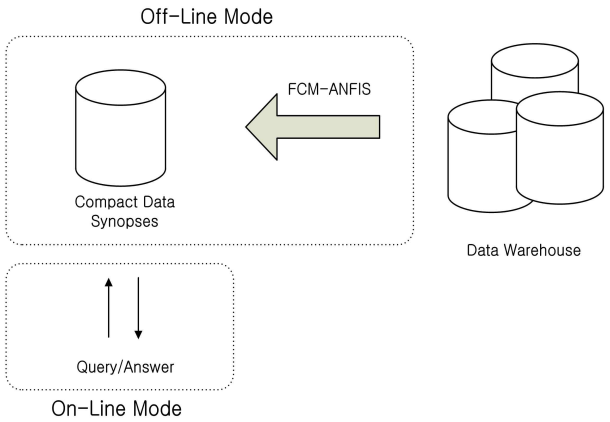
본 논문의 구성은 다음과 같다. 2장에서는 근사 질의응답의 개념과 기존 근사 질의응답 기법에 대하여 살펴본다. 3장에서는 FCM-ANFIS를 이용한 근사 질의응답 기법을 제안한고,. 4장에서 제안된 근사 질의응답 기법의 성능을 실험을 통하여 비교 분석한다. 마지막으로 5장에서 논문의 결론을 기술한다.

2. 관련연구

데이터웨어하우스에서 근사 질의응답은 집계 연산과 질의 수행의 가속화를 위해 사용되었다. 근사 질의응답의 기본 개념은 약간의 정보손실을 감수하면서 집계 계산과 질의 수행의 가속화에 초점을 두고 기초 데이터의 접근의 수를 최소화하거나 정확한 질의응답 계산의 회피에 의한 추정 결과를 제공하는 것이다[4]. 또는 실제 데이터에 대한 통계 요약정보를 사전에 계산하고 이 요약 정보를 기반으로 근사 응답을 제공하는 것이다[5]. 기존의 근사 질의응답 기법에는 샘플링 기반 기법[4, 6, 7, 16], 웨이블릿 기반 기법[8, 9, 10], 히스토그램 기반 기법[5, 14], 확률 및 정보 이론 기반 기법[3, 11, 12], 클러스터링 기반 기법[13, 15], 다항식 접근 기반 기법[2] 등의 방식들로 구분할 수 있다. NMF(Non-negative Multi- way array Factorization)[12] 접근법은 초기에 이차원 데이터를 분석하기 위해 개발되었으며 고차원 데이터를 쉽게 일반화 할 수 있다. 이 접근법은 데이터의 근사화를 제공하는 간단한 표현식을 찾음으로써 데이터 큐브의 상호작용과 패턴들을 발견할 수 있다. 발견된 패턴과 데이터 큐브의 각 셀에 확률을 할당하여 데이터 큐브의 근사적 모델을 생성한다. [13]은 연속 차원에서 근사질의를 위한 클러스터링 기법을 소개한다. 데이터 큐브의 새로운 압축된 표현으로 요구되는 저장 공간을 줄이고 집계질의에 대한 통계적 접근을 제안하여 다중 가우시안 확률 밀도 함수에 의해 각 클러스터들을 표현한다. [15]은 데이터큐브를 chunk들로 나누고 그 chunk를 K-Means 기법을 이용하여 클러스터링 한다. 균일 랜덤 샘플링은 전체 데이터베이스 R을 사전 계산된 균일 랜덤 샘플들의 열 S로 표현하는 것이다. 균일 랜덤 샘플링의 주요 장점은 사전처리가 단순하고 효과적이라는 것이고 단점은 거대한 데이터 분산의 통계적 문제이다[16]. 히스토그램 기법은 set-valued 질의들의 근사 응답을 제공하기 위해 [14]에서 사용되었다. 히스토그램은 작은 공간을 차지하며 추정 시간에 많은 오버헤드를 가지지 않고, 실생활에서 나타나는 치우진 분포를 정확하게 근사화하는데 알맞기 때문에 상업 데이터베이스 시스템에서 가장 광범위하게 사용되는 통계치이다. 버켓(bucket)은 그들 아이템의 합과 같은 어떤 집계 질의를 가지고 있으며, 버켓 분할 기법에는 같은 넓이의 범위를 갖는 equi-width histogram, 모든 버켓의 튜플들의 값이 같게 할당되는 equi-depth histogram 그리고 β-bucket $MaxDiff(V,A)$ histogram들이 있다. 히스토그램 기반 기법의 근사화의 정확성은 각 버켓으로 그룹화된 속성 값들에 의해 정의된다.

3. FCM-ANFIS을 이용한 근사 질의응답 기법

의사결정 시스템에서 사용자 질의에 대해 다량의 원 데이터를 통한 정확한 응답이 아닌 사전 계산된 데이터를 통해 근사 응답을 제공하기 위해 본 논문에서는 FCM 기법과 ANFIS 기법을 이용한 다차원 데이터 큐브의 근사 질의응답 기법을 제안한다. 다음 <그림 1>은 제안된 기법을 이용한 근사 질의응답 기법 프레임워크이다. 제안된 근사 질의응답 기법은 오프라인 모드(Off-Line Mode)와 온라인 모드(On-Line)로 나뉘며, 오프라인 모드에서 FCM-ANFIS를 이용하여 다차원 데이터 큐브의 지능형 모델을 생성한다. 이 생성된 모델은 온라인 모드에서 사용자의 질의에 응답하기 위해 사용된다.



[그림 1] Approximate query processing using FCM-ANFIS

패턴인식 분야에서 잘 알려진 Fuzzy C-Means (FCM)는 다른 클러스터링 기법과는 달리 데이터의 한 부분이 두 개 이상의 클러스터에 속하는 것을 허용하는 클러스터링 기법이다. 이 기법의 기본 개념은 다음과 같은 목적함수(objective function)를 최소화하는 클러스터링을 한다.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \|^2 \quad ,1 \le m < \infty \qquad (1)$$

$$subject\ to \sum_{i=1}^{c} u_{ik} = 1 \qquad (2)$$

식(1)에서 m은 퍼지화의 정도를 나타내는 상수로

서 1보다 큰 임의의 실수이고, u_{ij}는 클러스터 j에 대한 x_i의 소속도이며, x_i는 d-차원의 i번째 측정값이다. 또한 c_j는 j번째 클러스터의 d차원 중심이고, $\|*\|$는 측정 데이터와 중심 사이의 유사성을 표현하는 norm이다.

Adaptive Neuro-Fuzzy Inference System(ANFIS)는 언어적 입력 형태(rule)의 전제부와 1차 선형 방정식 형태의 결론부를 가지는 Takagi-Sugeno-Kang(TSK) 퍼지모델이라고도 한다[17]. TSK는 방정식의 차수에 따라 형태가 달라질 수 있으며 본 논문에서는 가장 일반적인 형태인 1차 방정식을 이용하였다. 결론부가 방정식 형태로 구성되어 있으며 비퍼지화 과정이 생략되고 주어진 실세계 문제가 방정식 형태로 구성되어 있어 접근하기 용이하다. 기본적인 TSK 퍼지 추론 시스템은 5개의 층을 가지며 2차원의 입력과 1차원의 출력을 가진다. 선형 공간으로 고려되는 결론부의 파라미터 학습은 선형 시스템에서 이용할 수 있는 다양한 기법을 통하여 최적화가 가능하므로 비선형 공간으로 고려되는 전제부의 학습에 좀 더 많은 관심을 가질 수 있다. 하지만 전제부의 구성 형태에 따라 모델의 크기가 달라질 수 있는데 일반적인 방법으로 격자 분할에 의한 소속 함수 생성의 경우 입력의 차원이 증가하거나 소속 함수가 증가하는 경우, 규칙의 수가 지수 함수적으로 증가한다. 이러한 문제점을 해결하기 위해여 규칙의 수가 단지 소속함수의 수에만 영향을 받는 클러스터링 기법에 의한 모델링이 도입되었다. 제안한 모델의 성능은 클러스터 개수의 선택에 영향을 받는다. 본 논문에서는 클러스터의 개수를 정하기 위해여 BIC기준을 사용한다.

제안된 기법에서 다차원 데이터 큐브에 대한 지능형 모델은 다음과 같은 단계를 수행하여 생성된다.

단계 1 : 입력 데이터 정규화

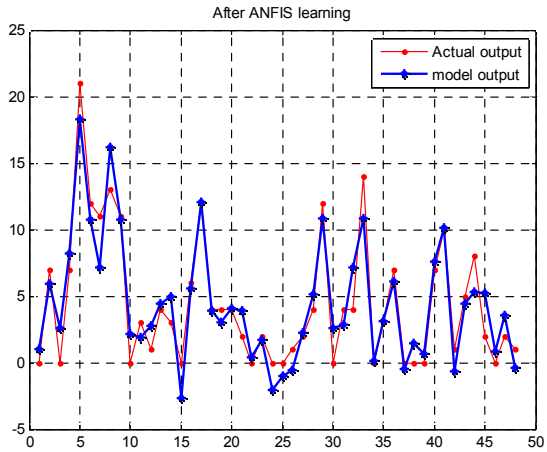
단계 2 : 정해진 클러스터 개수에 대해 FCM 실행 후 소속도와 중심값 생성

단계 3 : 데이터 범위내의 클러스터링 중심 확인

단계 4 : FCM 결과값을 이용하여 초기 FIS 생성

단계 5 : ANFIS 실행으로 학습을 통해 데이터에 적응된 모델 생성

다음 <그림 1>은 Government 데이터 테이블을 이용하여 지능형 모델을 생성한 후 추정된 값을 보여준다. 테이블의 데이터를 이용하여 FCM을 이용하여 클러스터링 한 후 ANFIS를 통한 결과이다.



[그림 1] ANFIS 결과

4. 실험 및 결과 분석

본 논문에서 제안한 FCM-ANFIS 기법을 이용한 근사 질의응답의 성능을 평가하기 위해 Microsoft SQL server에서 제공하는 FoodMart 데이터의 CUSTOMER 테이블을 이용한 질의응답 결과를 살펴본다. CUSTOMER 테이블은 1024개의 레코드를 가지는 5차원의 데이터 큐브를 생성한다. 이 테이블의 각 차원은 STATUS, INCOME, EDUCATION, CHILD, OCCUPATION으로 구성되어 있다. [12]에서 고려한 OLAP 연산 Slice, Dice, Roll-up의 종류로 구성되는 네 개의 질의를 통하여 FCM-ANFIS기법을 이용해 생성한 지능형 모델의 성능을 확인해 본다. 요약된 데이터 큐브를 통한 질의의 응답과 원 데이터 큐브를 통한 질의응답의 오차는 절대 오차를 이용하여 계산하며 다음 <표 1>과 같다. OLAP 연산에서 주로 행해지는 Slice, Dice, Roll-Up1에서는 기존의 기법보다 적은 오차를 가지는 것을 볼 수 있다. 따라서 제안된 FCM-ANFIS기법을 이용한 근사 질의응답 기법은 사용자의 질의에 대해여 정확성이 향상된 근사 응답을 제공함을 알 수 있다.

[표 1] Customer table에 대한 질의 결과

Query	Absolute Mean Error(E)	
	KR	NMF[12]
Slice	0.97	0.98
Dice	1.40	2.20
Roll-Up1	2.52	2.77
Roll-Up2	2.03	1.49

5. 결론

본 논문은 근사 질의응답을 위해 다차원 데이터큐브의 지능형 모델링을 제안하였다. FCM과 ANFIS를 이용하여 다차원 데이터 큐브의 모델을 생성하고 사용자의 질의에 대하여 원 데이터 큐브를 통해 응답을 하는 대신 지능형 모델을 통해 질의에 근사적으로 응답 한다. 실험 결과에서 대표적인 OLAP 연산에 대해 기존의 기법 보다 우수한 성능을 보임을 확인하였다. 따라서 제안된 FCM-ANFIS기법을 이용한 근사질의응답 기법은 사용자의 질의에 대해여 정확성이 향상된 근사 응답을 제공함을 알 수 있다. 향후 다차원 데이터 큐브의 갱신에 대한 연구가 필요하다.

사사

본 논문은 2008년도 지식경제부 성장동력기술 개발사업의 일환으로 (주)메타비즈의 위탁과제로 수행되었음

참고문헌

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kafmann Publishers, San Francisco, CA, USA, 2000.

[2] Alfredo Cuzzocrea, “Improving range-sum query evaluation on data cubes via polynomial approximation,” Data & Knowledge Engineering, vol. 56, Issue 2, pp.85-121, 2006.

[3] Themis Palpana and Nick Koudas, “Using Data cube Aggregates for Approximate Querying and Deviation Detection,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 11, November 2005.

[4] Phillip B. Gibbons and Yossi Matias, “New Sampling-Based Summary Statistics for Improving Approximate Query Answers,” Proceeding of the 1998 ACM Int. Conf. on Management of Data, pp.331-342, 1998.

[5] Viswanath Poosala and Venkatesh Ganti, “Fast approximate answers to aggregate queries on a data cube,” Eleventh International Conference on Scientific and Statistical Database Management, pp.24-33, 1999.

[6] Venkatesh Ganti, Mong Li Lee, and Ramakrishnan, "ICICLES: Self-tuning Samples for Approximate Query Answering," Proceedings of the 26th VLDB Conference, 2000.

[7] Brian Babcock, Surajit Chaudhuri, and Gautam Das, “Dynamic Sample Selection for Approximate Query Processing,” Proceedings of 22nd ACM SIGMOD International Conference, Management of Data (SIGMOD '03), pp. 539-550, 2003.

[8] Jeffrey Scott Vitter and Min Wang, "Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets," In Proceedings of the SIGMOD '99 Conference, pp. 193-204, 1999.

[9] Kaushik Chakrabarti, Minos Garofalakis, Rajeev Rastogi, and Kyuseok Shim, "Approximate Query Answering Using Wavelets," Proceedings of the 26th VLDB Conference, pp.111-122, 2000.

[10] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnam, and Martin J. Strauss, "Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries," Proceedings of the 27th VLDB Conference, 2001.

[11] Cyril Goutte, Rokia Missaoui and Ameur Boujenoui, "Data cube Approximation and Mining using Probabilistic Modelling," TR2007, NRC 2007.

[12] Rokia Missaoui, Cyril Goutte, Anicet K, Choupo and Ameur Boujenoui, "A Probabilistic Model for Data Cube Compression and Query Approximation," DOLAP 2007, ACM 10th International Workshop on Data Warehousing and OLAP, ACM Press, 2007.

[13] Jayavel Shanmugasundaram, Usama Fayyad, and P. S. Bradley, “Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions,” Proceeding of the 5th ACM SIGKDD international conference, ACM press, pp.223 - 232, 1999.

[14] Yannis E. Ionnidis and Viswanath Poosala, “Histrogram-Based Approximation of Set-Valued Query Answers,” 25th VLDB Conference, 1999.

[15] Feng Yu and Wang Shan, "Compressed data cube for Approximate OLAP Query Processing, " J. Comput. Sci. Technol., 17(5):625-635, 2002.

[16] Gautam Das, "Sampling Methods in Approximate Query Answering Systems", Invited Book Chapter, Encyclopedia of Data Warehousing and Mining. Editor John Wang, Information Science Publishing, 2005.

[17] J.-S. R. Jang, C.-T. Sum, and E, Mizutani, Neuro-Fuzzy and Soft Computing, Prentice Hall, 1997.