Journal of the Korea Academia-Industrial cooperation Society
Vol. 12, No. 8 pp. 3689-3696, 2011

http://dx.doi.org/10.5762/KAIS.2011.12.8.3689

지식 표현 방식을 이용한 근사 질의응답 기법

이선영[1], 이종연[1*]
[1]충북대학교 컴퓨터교육과

An Approximate Query Answering Method using a Knowledge Representation Approach

Sun Young Lee[1] and Jong Yun Lee[1*]
[1]**Department of Computer Education, Chungbuk National University**

요 약 의사결정 지원시스템에서 작업자들은 대량의 데이터 집계 연산을 요구하며, 데이터에 대한 정확한 응답보다는 경향 분석에 더 많은 관심을 가진다. 그러므로 정확한 응답보다 빠른 근사 질의응답을 제공하는 것이 필요하며 그것을 실현하기 위한 근사질의 응답 기법의 연구가 필요하다. 따라서 본 논문에서는 기존 연구들의 단점을 보안하고 근사 응답의 정확성을 향상시킬 수 있는 Fuzzy C-Means (FCM) 클러스터링 기반 Adaptive Neuro-Fuzzy Inference System (ANFIS)을 이용한 근사 질의응답 기법을 제안한다. FCM-ANFIS을 이용한 근사 질의응답 기법은 다차원 데이터의 지식 표현 모델을 생성함으로써 거대한 다차원 데이터 큐브에 직접적인 접근 없이 집계 질의 수행이 가능하다. 비교실험을 통하여 제안된 기법이 기존의 NMF 기법보다 근사 질의응답의 정확성이 향상되었음을 확인한다.

Abstract In decision support system, knowledge workers require aggregation operations of the large data and are more interested in the trend analysis rather than in the punctual analysis. Therefore, it is necessary to provide fast approximate answers rather than exact answers, and to research approximate query answering techniques. In this paper, we propose a new approximation query answering method which is based on Fuzzy C-means clustering (FCM) method and Adaptive Neuro-Fuzzy Inference System (ANFIS). The proposed method using FCM-ANFIS can compute aggregate queries without accessing massive multidimensional data cube by producing the *KR* model of multidimensional data cube. In our experiments, we show that our method using the *KR* model outperforms the NMF method.

Key Words : Data cube, Approximate query answering, FCM-ANFIS, Data warehouse

1. 서 론

데이터 웨어하우스는 전문가, 매니저, 분석가들이 더 빠르고 더 나은 결정을 만드는 것을 돕기 위해 구축된 대규모 집계 데이터의 집합이며, 의사결정 지원시스템(Decision Support System, DSS)에서 가치 있고 유용한 지식을 추출하기 위해 사용된다[1,2]. 각 시스템들은 고차원의 데이터를 가지며 이 데이터의 크기와 함께 차원의 수는 DSS 기반 응용들의 효과성에 영향을 주는 주요한 요소들 중 하나이다. 이런 데이터 웨어하우스는 보통 다차원 데이터베이스 구조로 모델링되며, 각 차원(dimension)은 스키마에 있는 하나의 속성이나 속성의 집합에 해당하고 배열의 축에 대응된다. count나 sales등과 같은 집계측도(measure)들은 배열의 각 셀(cell)에 저장되며 데이터 분석의 대상이 된다.

1.1 연구동기

의사 결정 시스템의 응용들은 보통 데이터의 탐사를 위주로 하여 집계 질의를 요구하지만 데이터들에 대한 꼼꼼한 분석보다 경향분석에 더 많은 관심을 가지며, 정량적보다 정성적인 분석질의로 소수점 이하의 정답을 필요로 하지 않는다[3,4]. 따라서 정확한 응답보다는 빠른

*교신저자 : 이종연(jongyun@chungbuk.ac.kr)
접수일 11년 06월 14일 수정일 (1차 11년 07월 06일, 2차 11년 07월 19일) 게재확정일 11년 08월 11일

근사 질의응답을 제공하는 것이 필요하며 그것을 실현하기 위해서 실제 값에 가까우면서 빠른 시간 내에 질의응답을 구하는 근사 질의응답(Approximate Query Answering: AQA)이 필요하다. 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[5]. 또는 실제 데이터에 대한 통계 요약정보를 사전에 계산하고 이 요약 정보를 기반으로 근사 응답을 제공한다[6].

최근 다양한 근사 질의응답 기법의 연구가 진행되었다. 샘플링 기반 기법(sampling-based techniques) [5,7,8,17,21], 웨이블릿 기반 접근법(wavelet-based approach)[9-12], 히스토그램 기반 기법(histogram-based techniques) [6,13], 확률 및 정보 이론(probability and information theory) 기반 기법[3,4,14,15], 클러스터링 기반 기법[2,16] 등이 있다. 기존 AQA 방법들은 주로 데이터 큐브를 압축하고, 압축된 큐브를 기반으로 AQA를 제공하는 방식을 채택하고 있다[3]. 본 논문에서는 의사 결정 지원 시스템에서 기존 AQA 기법들보다 정확성이 우수하며 사전 계산된 데이터의 저장이 적은 기법을 제안한다.

1.2 기여도

퍼지 개념을 도입한 Fuzzy C-Means(FCM) 클러스터링 기법[22]과 Adaptive Neuro-Fuzzy Inference System(ANFIS) [22-25]을 이용해 다차원 데이터 큐브의 지식 표현(knowledge representation)을 생성하면 거대한 다차원 데이터 큐브의 직접적인 접근 없이 집계 질의를 수행 할 수 있다. 따라서 본 논문에서는 근사 질의응답을 효율적으로 처리 할 수 있는 FCM- ANFIS을 이용한 근사 질의응답 기법을 제안한다. FCM-ANFIS을 이용하여 학습된 KR 모델은 다차원 데이터 큐브의 데이터 특성을 가지며 질의에 대한 근사 응답을 제공할 수 있는 데이터 큐브의 지식 표현이 된다. 이 제안된 기법은 적은 수의 파라미터로 데이터를 표현함으로써 다차원 데이터 큐브의 압축된 표현을 위한 저장 공간과 응답 시간을 줄일 수 있다. 또한 기존의 기법에 비해 근사 응답의 정확도를 향상시킬 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 기존 근사 질의응답 기법에 대하여 살펴보고 3장에서는 FCM-ANFIS를 이용한 근사 질의응답 프레임워크와 근사 질의응답 알고리즘을 제안한다. 4장에서는 제안된 근사 질의응답 기법의 성능을 실험하고 결과를 분석한다. 마지막으로 5장에서 본 논문을 간략히 요약하고 결론을 기술한다.

2. 관련 연구

데이터 웨어하우스에서 근사 질의응답은 집계 연산과 질의 수행의 가속화를 위해 사용되며 온라인 집계 방법과 사전 계산 방법으로 나눌 수 있다. 온라인 집계 방법은 질의 시간에 샘플링을 실행하여 응답에 연속적인 정제를 실행하며 사용자는 무엇을 재정의하고 언제 멈추는지를 제어 할 수 있다[19]. 사전 계산 방법은 오프라인동안 데이터 웨어하우스에 대한 사전 계산을 통해 요약 데이터를 생성하고 온라인 시 요약 데이터를 이용해 질의에 응답한다[18].

확률 모델 기반 기법은 가능한 적은 파라미터들을 사용하여 데이터 큐브의 데이터들에 확률을 할당하고 근사적 표현을 생성하는 것이 목적이다[14,15]. NMF(non-negative multi-way array factorization) 접근법은 초기에 이차원 데이터를 분석하기 위해 개발되었으며 데이터 큐브의 분석과 탐사에 적용하면서 데이터의 근사화를 제공하는 간단한 표현식(concise representation)을 찾음으로써 데이터 큐브의 상호작용과 패턴들을 발견할 수 있다. 발견된 패턴과 데이터 큐브의 각 셀에 확률을 할당하여 확률 모델을 생성한다. 가능한 적은 파라미터를 이용해 데이터 큐브를 근사화하려고 하며 적은 차원에서 가장 큰 이익을 가진다[15]. Palpana et al.[4]는 information entropy을 기반으로 저장된 집계 값들을 사용하여 근사적으로 질의에 응답할 수 있는 기법을 제안하였다. 요약된 형태의 데이터로부터 원래의 데이터를 재구성하기 위해 정보 이론 원리를 사용한다.

Shanmugasundaram et al.[16]는 연속 차원에서 근사질의를 위한 클러스터링 기법을 소개하였으며, Yu et al.[2]는 chunk들로 데이터 큐브를 나누고 그 chunk들을 가지고 K-means 클러스터링 기법을 이용하여 데이터 큐브의 압축된 표현을 생성한다. Gibbons et al. [5]는 전체 데이터베이스 R을 사전 계산된 균일 랜덤 샘플들의 열 S로 표현하여 원 데이터에 대한 좀 더 작은 표현을 달성할 수 있게 해주는 샘플링기반 기법을 제안하였다. 빠른 근사 응답을 제공하기 위해 S에 대해 간단히 질의를 실행하고 결과를 비율에 따라 조정한다.

그 외에 웨이블릿 기반 접근법[9-12], 히스토그램 기반 기법[6,13], DCT기반 기법[20] 등이 있다.

3. FCM-ANFIS 기반 근사 질의응답

3.1 FCM 클러스터링과 ANFIS 모델

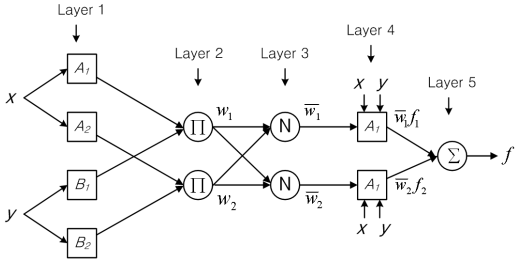
클러스터링은 같은 클래스 내의 유사성은 높이고, 다른 클러스터 클래스 간의 이질성을 높게 데이터들을 묶는다. 패턴인식 분야에서 잘 알려진 Fuzzy C-Means (FCM)[22]는 fuzzy ISODTA라고도 하며 다른 클러스터링 기법과는 달리 데이터의 한 부분이 두 개 이상의 클러스터에 속하는 것을 허용하며 멤버쉽 등급에 따라 데이터를 분류하는 클러스터링 알고리즘이다[22]. 이 기법의 기본 개념은 다음과 같은 목적함수(objective function)를 최소화하는 클러스터링을 한다.

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

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

식(1)에서 m은 퍼지화의 정도를 나타내는 상수로서 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[22,23].



[그림 1] ANFIS architecture
[Fig. 1] ANFIS architecture

그림 1은 1차 Sugeno 퍼지 모델의 ANFIS를 보여준다. TSK는 방정식의 차수에 따라 형태가 달라질 수 있으며 본 논문에서는 가장 일반적인 형태인 1차 방정식을 이용하였다. 두 개의 입력 x, y와 하나의 출력 f을 가지는 TSK 퍼지 추론 시스템을 이용한다. 각 층의 특성과 학습 절차는 다음과 같다[22,23].

Layer 1 : 이 층의 노드 i는 노드 함수를 가지는 적응 노드이다.

$$O_{1,i} = \mu_{A_i}(x), \text{ for } i = 1, 2 \text{ or}$$
$$O_{1,i} = \mu_{B_{i-2}}(x), \text{ for } i = 3, 4. \qquad (3)$$

여기서 x(또는 y)는 노드 i의 입력이고 A_i(또는 B_{i-2})는 이 노드에 할당된 언어적 변수(linguistic label)이다. A에 대한 소속함수는 다양한 소속함수를 사용할 수 있다. 본 논문에서 는 다음 식(4)과 같은 가우시안 소속함수를 이용하였다.

$$\mu_{A_i}(x) = e^{-\frac{1}{2}\left(\frac{x-c_i}{\sigma_i}\right)^2} \qquad (4)$$

여기서 $\{\sigma_i,\ c_i\}$는 전제부 파라미터이며, σ_i는 소속함수의 폭이고, c_i는 소속함수의 중심이다.

Layer 2 : 이 층의 각 노드는 모든 입력 신호의 곱을 출력하는 Π로 표현되는 노드이다.

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(y), \text{ for } i = 1,2 \qquad (5)$$

Layer 3 : 이 층의 i번째 노드는 모든 규칙의 활성도의 합에 대한 i번째 규칙의 활성강도의 비를 계산하며, N로 표현된다. 즉 이 층의 출력은 정규화된 값을 출력한다.

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1 + w_2},\ i = 1,2 \qquad (6)$$

Layer 4 : 이 층의 모든 노드는 정규화 된 가중치를 결론부의 파라미터와 곱하여 출력한다.

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i (p_i x + q_i y + r_i),\ i = 1,2 \qquad (7)$$

여기서 $\overline{w}_i$는 계층3의 출력인 정규화된 활성강도이고 $\{p_i,\ q_i,\ r_i\}$는 이 노드의 파라미터 집합이다.

Layer 5 : 이 계층의 단일 노드는 Σ로 표현되며, 가중평균법(weighted average)에 의한 최종 출력을 얻는다.

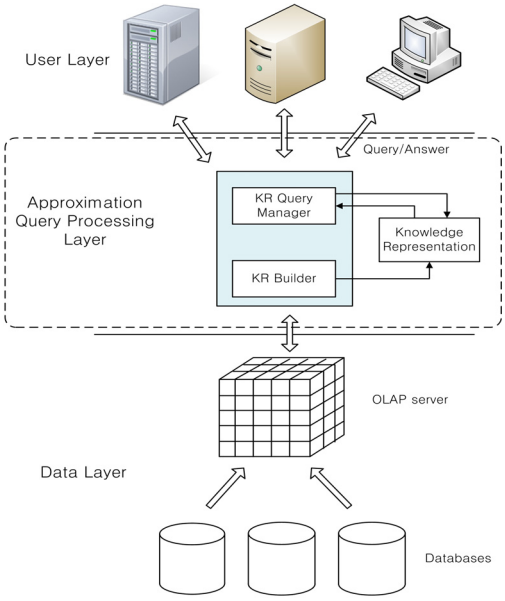
$$O_{5,i} = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i},\ i = 1,2 \qquad (8)$$

전제부와 결론부의 파리미터를 추정하기 위해 혼합 학습방법을 사용한다. 전방향 학습에서 전제부 파라미터를 고정시키고 결론부 파라미터를 추정하고 최적의 결론부 파라미터를 구하면 역방향 학습이 실행된다. 전방향 학습에서는 least squared method가 적용되며 후방향 학습에는 gradient descent method가 적용된다. ANFIS에 사용된 혼합 학습 방법은 [22]에서 자세히 논하고 있다.

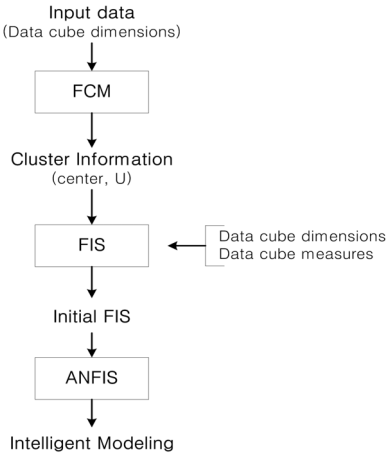
TSK 퍼지 모델은 언어적 입력을 가지며 출력은 1차 선형 방정식 형태를 가진다. 이 경우 결론부의 최종 출력을 구하기 위하여 필요한 비퍼지화 과정이 불필요하며, 비선형 입력 공간을 뉴로-퍼지 규칙을 통하여 규칙의 수만큼의 부분선형 공간으로 결론부를 구성하는 형태의 비선형 공간을 선형 공간으로 변환하는 함수로 고려될 수 있다. 또한 선형 공간으로 고려되는 결론부의 파라미터 학습은 선형 시스템에서 이용할 수 있는 다양한 기법을 통하여 최적화가 가능하므로 비선형 공간으로 고려되는 전제부의 학습에 좀 더 많은 관심을 가질 수 있다. 하지만 전제부의 구성형태에 따라 모델의 크기가 달라질 수 있는데 일반적인 방법으로 격자분할에 의한 소속함수 생성의 경우 입력의 차원이 증가하거나 소속함수가 증가하는 경우 규칙의 수가 지수 함수적으로 증가한다. 이러한 문제점을 해결하기 위하여 규칙의 수가 단지 소속함수의 수에만 영향을 받는 FCM 클러스터링 기법에 의한 모델링을 적용한다.

3.2 FCM-ANFIS을 이용한 근사 질의응답

본 절에서는 의사결정 시스템에서 사용자 질의에 대해 다량의 원 데이터를 통한 정확한 응답이 아닌 사전 계산된 데이터를 통해 근사 응답을 제공하기 위해 FCM-ANFIS기법을 이용한 다차원 데이터 큐브의 근사 질의응답을 제안한다. 그림 2는 세 계층을 갖는 근사 질의응답 시스템의 구조를 보여준다. *User Layer*는 사용자 계층으로 전문가, 매니저, 분석가들이나 응용 시스템이 OLAP을 수행한다. *Data Layer*는 OLAP 서버가 위치한다. *Approximation Query Processing Layer*는 FCM-ANFIS를 이용하여 *KR*(*Knowledge Representation*)을 생성한다. *KR Query Manger*는 *User Layer*를 통해 들어오는 질의를 *KR*에 알맞게 변형하여 질의하고, 질의에 대한 응답을 사용자에게 전달해 주는 역할을 한다. *KR Builder*는 OLAP 서버의 다차원 데이터 큐브를 FCM-ANFIS을 이용한 지능형 모델링을 통해 새로운 지식 표현을 생성한다. 즉, 제안된 근사 질의응답 기법은 오프라인 모드(Off-Line Mode)에서 *KR Builder*가 FCM- ANFIS를 이용하여 다차원 데이터 큐브의 지식 표현(*KR*)을 생성하고, 온라인 모드(On-Line)에서는 생성된 *KR* 모델을 이용하여 *KR Query Manger*가 사용자의 질의에 응답한다.



[그림 2] 근사 질의응답 시스템의 3계층 지식표현 구조
[Fig. 2] A three-tiered knowledge representation architecture of approximate query answering system



[그림 3] FCM-ANFIS을 이용한 지식 표현 생성과정
[Fig. 3] Generation process of the KR model using FCM-ANFIS

*KR Builder*에서 OLAP server의 다차원 데이터 큐브에 대한 *KR* 생성은 그림 3과 같은 과정을 수행한다. 이 과정은 그림 4의 *buildKR* procedure로 다시 표현할 수 있다. 그림 4의 단계 2에서 입력 데이터를 정규화하고 단계 3에서 정해진 클러스터 개수에 대해 FCM 실행 후 소속도와 중심값을 생성한다. 생성한 중심값을 데이터 범위내인지 확인 후 단계 4에서 FCM 결과값을 이용하여 초기 FIS 생성한다. 마지막으로 단계 5에서 ANFIS를 실행하여 학습을 통해 데이터에 적응된 지능형 모델(*KR*)을 생성한다.

Procedure 1 *buildKR(C, k)*

Input The data cube C; the number of clusters k;
Output The knowledge representation of C, *KR*;

```
1: Begin
2:    Normalize an input data cube C ;
3:    [centers, U] = FCM(k, C);    // Clustering
4:    f = FIS(centers, U);         // Generate a fuzzy
                                   //      inference system
5:    KR = ANFIS(f, C);            // Build the KR model
6: End;
```

[그림 4] The *buildKR* procedure
[Fig. 4] The *buildKR* procedure

3.3 KR 모델을 이용한 근사 질의응답 알고리즘

n-차원 데이터 큐브를 $C=\{D_1, D_2, ..., D_n; M\}$라고 하자. $D_1, D_2, ..., D_n$은 n개의 차원을 나타내며, M은 데이터 큐브의 측정값을 나타낸다. 데이터 큐브를 *KR* 모델로 변환할 때 각 셀이 갖고 있는 정확한 정보들은 잃을 수 있지만 *KR* 모델은 각 셀의 값을 추정할 수 있다. 정통적인 OLAP 질의들은 roll-up, drill-down, slice, and dice 등이 있으며 이 연산들은 영역-합 질의에 기반 한다. RA_i를 차원 i에 대한 특정 구간을 나타낸다고 하면 $RA = RA_1 \times RA_2 \times \cdots \times RA_n$의 영역을 가지는 집계 질의를 나타낼 수 있다. 다음 그림 5은 RA의 범위 질의에 대한 *KR* 모델의 근사 질의응답에 대한 알고리즘이다. 범위 RA의 질의가 입력으로 들어오면 근사 질의응답(AQA) 알고리즘은 각 범위 RA_i에 대하여 단계 3을 반복한다. 단계 3에서 *KR* 모델을 이용하여 각 셀에 대응하는 값을 찾아낸다. 단계 5에서는 단계 3에서 구한 각 차원에 해당하는 값을 가지고 집계 연산을 실행하여 근사 응답 $\tilde{A}$를 구한다.

Algorithm 1 AQA algorithm

Input query RA;
Output Approximate query answer $\tilde{A}$;

```
1: Begin
2:    for each range RA_i
3:       Find the corresponding values for every
         cell on KR model.
4:    end for
5:    Compute the aggregation Ã;
6: End;
```

[그림 5] KR 모델을 이용한 AQA 알고리즘
[Fig. 5] AQA algorithm using *KR* model

4. 실험 및 결과 분석

4.1 실험 데이터

본 논문에서 제안한 FCM-ANFIS을 이용한 근사 질의응답 기법의 정확성 및 효율성을 확인하기 위해 두 종류의 데이터 집합을 사용하였다.

첫 번째 Governance[14,15] 데이터 테이블은 주식시장에 상장된 214개의 캐나다 회사 샘플의 4차원의 큐브로 USSX, Duality, Size, QI의 차원을 가지며 집계측도는 상장 회사의 수이다. 표 1는 4차원에 따른 회사의 수를 제공하는 fact 테이블이다.

[표 1] Governance 데이터 큐브
[Table 1] Data cube for the Governance example

		DUALITY : No			Yes		
USSX	SIZE	QI:Low	Med	High	Low	Med	High
NO	1	0	7	0	4	3	0
	2	7	21	12	6	12	4
	3	11	13	11	4	4	2
	4	0	3	1	0	2	0
Yes	1	0	1	2	0	0	0
	2	4	12	0	7	10	1
	3	4	4	14	5	8	2
	4	0	3	7	0	2	1

두 번째 데이터 집합은 Microsoft SQL server에서 제공하는 FoodMart의 CUSTOMER[15] 테이블이다. 이 테이블은 10281개의 레코드를 가지는 5차원의 데이터 큐브를 생성한다. 테이블의 각 차원은 STATUS(marital_status, 2), CHILD(total_ children, 6), INCOME(yearly_income, 8), EDUCATION(education, 5), OCCUPATION(occupation, 5)으로 구성되어 있다. 괄호안의 이름은 실제 테이블의 이름이며, 숫자는 각 차원의 속성값의 수를 나타낸다. 표 2는 실험에 사용한 두 개 데이터의 특징을 요약한 것이다. 각 데이터의 차원 수, 밀도와 영이 아닌 셀의 수를 나

타낸다.

[표 2] 데이터 큐브들의 특성
[Table 2] Data cubes characteristics

Data Cube	Dimension	Size	Non zero cells	Sparsity [%]
Governance	4	48	35	72.91
Customer	5	2400	877	36.54

4.2 근사 질의응답 결과 분석

첫 번째 실험은 Governance 데이터[14,15]를 이용하여 NMF[14, 15]모델과 비교하였다. 데이터 큐브 근사화에 대한 *KR* 모델의 장점을 보여주기 위해 다음과 같은 질의를 사용하였다. 첫 번째 질의는 *KR* 모델과 NMF를 이용하여 원 데이터를 추정하는 것이고, 나머지 두 질의는 차원 선택과 roll-up 연산을 보여준다[14].

1. Substitution: Counts of firms according to SIZE, USSX, QI and DUALITY(the original cube).
2. Selection: Counts according to SIZE and USSX, for firms with low governance quality(QI=Low), where the CEO is not also chairman of the board(DUALITY = No).
3. Aggregation: Counts of firms according to SIZE and USSX, aggregated over all other variables.

KR 모델을 이용한 첫 번째 질의의 결과는 표 3에 나타내었다. 원 데이터 표 1과 *KR* 모델에 의해 추정된 값 사이의 절대평균오차는 0.35이다.

[표 3] *Governance* 데이터의 추정결과
[Table 3] Estimated data cube for the *Governance*

		DUALITY : No			Yes		
USSX	SIZE	QI:Low	Med	High	Low	Med	High
NO	1	1	7	0	2	3	0
	2	7	21	12	6	12	4
	3	11	13	11	4	4	2
	4	2	2	2	0	1	0
Yes	1	0	1	1	0	1	1
	2	4	12	0	6	11	2
	3	4	4	14	5	8	2
	4	0	2	6	0	3	1

두 번째 질의는 몇몇 차원의 특정 값들에 대한 선택 질의로 그림 6에 결과를 나타내었다. 그림 6(a)는 원 데이터에 대한 두 번째 질의 결과를 보여주고 그림 6(b)와 그림 6(c)는 각각 *KR* 모델과 NMF에 대한 질의 결과를 보여준다. NMF의 절대평균오차는 1.625이지만, *KR* 모델의 절대평균오차는 0.375이다.

(Data)	USSX	
SIZE	No	Yes
1	0	0
2	7	4
3	11	4
4	0	0

(a)

(*KR*)	USSX	
SIZE	No	Yes
1	1	0
2	7	4
3	11	4
4	2	0

(b)

(NMF)	USSX	
SIZE	No	Yes
1	2	0
2	11	5
3	7	3
4	1	0

(c)

[그림 6] 선택 질의(SIZE×USSX for QI=Low and DUALITY=No) 결과: (a) 원 데이터; (b) *KR* 모델 결과; (c) NMF 결과.
[Fig. 6] Results from the selection query SIZE×USSX for QI=Low and DUALITY=No: (a) Results on original data; (b) Results on the *KR* model; (c) Results on NMF.

마지막 질의는 집계질의이다. *KR* 모델로부터 전체 데이터 큐브의 값을 추정하고 집계를 실행한다. 그림 7에 결과를 나타내었다. 두 모델의 오차는 0.675이다. 이 데이터를 이용한 실험에서 우리의 제안된 기법은 NMF와 비슷하거나 더 나은 결과를 갖는다.

(Data)	USSX	
SIZE	No	Yes
1	14	3
2	62	34
3	45	37
4	6	13

(a)

(*KR*)	USSX	
SIZE	No	Yes
1	13	4
2	62	35
3	45	37
4	7	12

(b)

(NMF)	USSX	
SIZE	No	Yes
1	13	2
2	64	34
3	44	37
4	6	13

(c)

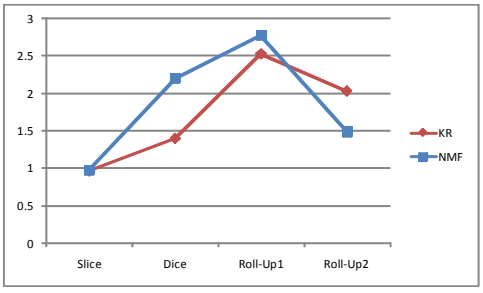
[그림 7] SIZE와 USSX에 따른 집계질의 결과 (a) 원 데이터 결과; (b)*KR* 모델 결과; (c)NMF 결과.
[Fig. 7] Results from the distribution aggregation query of firms according to SIZE and USSX: (a) Results on the original data cube; (b) Results on the *KR* model; and (c) Results on NMF.

두 번째 실험은 [15]에서 고려한 질의 네 가지를 본 논문에서도 Customer 테이블에 적용하여 제안한 기법의 정확성을 확인해 보았다. 질의는 OLAP의 주요 연산인 slice, dice, Roll-up으로 다음과 같다.

1. Slice : Number of customers according to STATUS, INCOME, CHILD, and OCCUPATION for customers with EDUCATION=4.
2. Dice : Number of customers according to STATUS, INCOME and OCCUPATION for customers with EDUCATION=4.

3. Roll-Up1 : Number of customers according to INCOME, OCCUPATION and EDUCATION.
4. Roll-Up2 : Number of customers according to the five dimensions (STATUS, INCOME, CHILD, OCCUPATION and EDUCATION).

FCM-ANFIS을 이용해 생성한 지능형 모델(*KR*)을 통한 질의응답과 원래 데이터 큐브를 통한 질의응답의 오차는 절대 오차를 이용하여 계산한다. 다음 그림 8은 각 질의에 대해 제안된 기법 *KR* 모델과 NMF[14,15]기법에 따fms 오차를 보여준다. 질의 4번 Roll-up2 질의의 결과를 제외한 나머지 결과에 대해 제안된 기법이 더 적은 오차를 가지는 것을 알 수 있다. 따라서 FCM-ANFIS을 이용한 근사 질의응답 기법이 NMF 기법에 비해 적은 오차를 가지며, 더 정확한 근사 응답을 제공하는 것을 알 수 있다.



[그림 8] KR과 NMF의 질의에 대한 오차
[Fig. 8] Error for query on KR and NMF

표 4는 각 데이터에 대한 *KR* 모델과 NMF의 압축률을 보여준다. *KR*은 NMF 보다 적거나 비슷한 파라미터를 가지지만은 위의 결과들에서 보여주는 것과 같이 더 높은 정확성을 갖는다.

[표 4] *KR* 모델과 NMF의 압축률
[Table 4] The compression rate of the KR model and NMF

Data Cube	Model	Rules	Parameters(f)	CR[a)][%]
Governance N[b)]=48	KR	4	20	58.3
	NMF	·	24	50.0
Customer $N=2400$	KR	18	108	95.5
	NMF	·	110	95.4

a) compression rate = $1-\frac{f}{N}$, b) total number of cell

5. 결론 및 향후연구

본 논문은 의사 결정 지원시스템의 빠른 근사 응답을 제공하기 위해 데이터 큐브의 압축된 표현을 제공하였다. FCM-ANFIS을 이용하여 학습된 지식 표현(*KR*) 모델은 다차원 데이터 큐브의 특성을 가지며 거대한 다차원 데이터 큐브의 직접적인 접근 없이 질의를 수행 할 수 있었다. 제안된 *KR* 모델을 이용한 근사 질의응답 기법의 성능이 기존의 제안된 NMF 보다 더 정확한 근사 응답을 제공함을 실험을 통하여 보였다. 이 제안된 기법의 장점은 다음과 같다. 첫째, 적은 파라미터로 데이터를 표현함으로써 다차원 데이터 큐브의 압축된 표현을 위한 저장 공간을 줄일 수 있다. 둘째, 원 다차원 데이터 큐브에 비해 적은 파라미터를 접근함으로써 질의 응답 시간을 줄일 수 있다. 셋째, 근사 응답의 정확도를 향상시킬 수 있다.

향후, 다차원 데이터 큐브에 대한 KR 모델의 설계 최적화와 오프라인 동안의 유지보수에 대한 연구가 필요하다.

References

[1] J. Han et al., *Data Mining: Concepts and Techniques*, Morgan Kafmann Publishers, 2000.

[2] F. Yu et al., "Compressed data cube for Approximate OLAP Query Processing," Journal of Computer Science and Technology, Vol. 17, Issue 5, pp.625-635, 2002.

[3] A. Cuzzocrea, “Overcoming Limitations of Approximate query Answering in OLAP,” Proceedings of the 9th International Database Engineering & Application Symposium (IDEAS'05), pp.200-209, 2005.

[4] T. Palpana et al., “Using Data cube Aggregates for Approximate Querying and Deviation Detection,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 11, 2005.

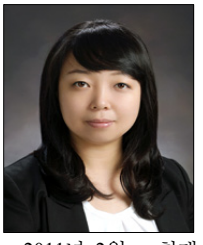
[5] P. B. Gibbons et al., “New Sampling-Based Summary Statistics for Improving Approximate Query Answers,” Proceeding of the 1998 ACM Int. Conf. on Management of Data, pp. 331-342, 1998.

[6] V. Poosala et al., “Fast approximate answers to aggregate queries on a data cube,” Eleventh International Conference on Scientific and Statistical Database Management, pp.24-33, 1999.

[7] V. Ganti et al., "ICICLES: Self-tuning Samples for Approximate Query Answering," Proceedings of the

26th VLDB Conference, Cairo, Egypt, 2000.
[8] R. Jin et al., “New Sampling-Based Estimators for OLAP Queries,” The 22nd International Confernece on Data Engineering, ICDE'06, 2006.
[9] J. S. Vitter et al., “Data Cube Approximation and Histograms via Wavelets,” Proceedings of Seventh International Conference on Information and Knowledge Management (CIKM'98), Washington D.C., November 1998.
[10] J. S. Vitter et al., "Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets," In Proceedings of the SIGMOD '99 Conference, pages 193-204, 1999.
[11] K. Chakrabarti et al., "Approximate Query Answering Using Wavelets", Proceedings of the 26th VLDB Conference, Cairo, Egypt, pages 111-122, 2000.
[12] A. C. Gilbert et al., "Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries", Proceedings of the 27th VLDB Conference, Romma, Italy, 2001.
[13] Y.E. Ionnidis et al., “Histogram-Based Approximation of Set-Valued Query Answers,” 25th VLDB Conference, 1999.
[14] C. Goutte et al., "Data cube Approximation and Mining using Probabilistic Modelling," TR 2007, NRC 2007.
[15] R. Missaoui et al., "A Probabilistic Model for Data Cube Compression and Query Approximation," DOLAP 2007, ACM 10th International Workshop on Data Warehousing and OLAP, ACM Press, 2007.
[16] J. Shanmugasundaram et al., “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.
[17] B. Babcock et al., “Dynamic Sample Selection for Approximate Query Processing,” Proceedings of 22nd ACM SIGMOD International Conference, Management of Data (SIGMOD '03), pp. 539-550, 2003.
[18] S. Acharya et al., “The Aqua Approximate Query Answering System,” SIGMOD, 1999.
[19] J. M. Hellerstein et al., “Online Aggregation,” Proceedings of ACM SIGMOD Conference, 1996.
[20] Wen-Chi Hou, Cheng Luo, Zhewei Jiang, and Feng Yan, "Approximate Rang-sum queries over data cubes using cosine transform," 2008.
[21] 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.
[22] J.S.R. Jang et al., *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and machine Intelligence,* Prentice Hall, 1997.
[23] J.S.R. Jang, "ANFIS: Adaptive network-based fuzzy inference system," IEEE Transactions on Systems, Man and Cybernetics, Vol. 23 (3) pp. 665-685, 1993.
[24J .S.R. Jang, "Input selection for ANFIS learning," in: Proceedings of the Fifth IEEE International Conference on Fuzzy Syste, pp.1493-1499, 1996.
[25] M. A. Denai, et al., "ANFIS based modelling and control of non-linear systems: a tutorial," IEEE International Conference on Systems, Man and Cybernetics, pp.3433-3438, 2004.

이 선 영(Sun Young Lee) **[정회원]**



- 2001년 2월 : 충북대학교 전기공학과(학사)
- 2005년 8월 : 충북대학교 전기공학과(공학석사)
- 2010년 8월 : 충북대학교 컴퓨터교육과(교육학박사)
- 2010년 6월 ~ 2011년 1월 : 한국생명공학연구원 박사후연구원
- 2011년 2월 ~ 현재 : 한국한의학연구원 박사후연구원

<관심분야>
데이터베이스, Bioinformatics

이 종 연(Jong Yun Lee) **[정회원]**

- 1985년 : 충북대학교 전자계산기공학과(공학사)
- 1987년 : 충북대학교 전자계산기공학과(공학석사)
- 1999년 : 충북대학교 전자계산학과(이학박사)
- 1999년 3월 ~ 2003년 2월 : 강원대학교 삼척캠퍼스 정보통신공학과 조교수
- 2003년 3월 ~ 현재 : 충북대학교 컴퓨터교육과 교수
- 2010년 3월 ~ 현재 : 한국컴퓨터교육학회 이사(현)
- 2010년 5월 ~ 현재 : 한국융합학회장(현)

<관심분야>
질의처리 및 최적화, 근사질의응답(AQA), 데이터베이스