# CEP 를 위한 데이터 마이닝 기법 연구

강동현, 황부현
전남대학교 전자컴퓨터공학과
e-mail : procsia@naver.com

# A Study of Data Mining Techniques for CEP

Donghyun Kang, Buhyun Hwang
Dept. of Electronics and Computer Engineering, Chonnam National University

요 약

최근에 이슈가 되고 있는 빅 데이터 처리 방법중의 하나로 CEP 가 있다. 그러나 CEP 는 사전에 정의된 질의에 해당되는 이벤트만을 선별하여 패턴 매칭 등의 기능을 수행하므로, 새로이 발견되는 이벤트를 찾는데 제약이 있다. 또한 실시간으로 생산되는 빅 데이터에 기초한 다양한 패턴 탐사에 한계를 노출하고 있다. 이 논문에서는, CEP 환경에서 빅 데이터 사이에 존재하는 다양한 이벤트와 패턴 탐사를 위한 실시간 데이터 마이닝 기법을 제안한다. 제안 방법은 CEP 엔진을 위한 고급의 패턴 매칭을 개발하고, CEP 를 위한 실시간 데이터 마이닝 기법을 개발한다. 마지막으로, 기존의 CQL 을 확장하여 개발한다. 이러한 방법을 통하여 기존의 CEP 의 기능적인 한계를 극복할 수 있다.

## 1. 서론

IT 기술의 발전으로 인하여 멀티미디어 및 센서 네트워크의 다양한 응용 프로그램으로부터 생산되는 데이터 양이 기하급수적으로 늘어나게 되었다. 빅 데이터(Big data)란, 기존의 트랜잭션에서 처리하는 데이터 규모가 커진 것뿐만 다양한 응용 프로그램으로부터 발생한 방대한 범위와 크기를 갖는 데이터를 의미한다[1]. 특히 스마트 폰 관련 기술이 급격히 발전함으로 인하여 소셜 네트워크 서비스(SNS)와 같은 네트워크 기반 애플리케이션이 발달하게 되었다. 이는 전통적인 데이터베이스 시스템에서 발생되는 데이터에 더하여 다양한 형태의 텍스트, 이미지, 동영상과 같은 비정형 데이터의 폭발적인 증가를 발생시켰으며, 효율적으로 데이터를 처리하는 일이 매우 중요한 연구분야가 되고 있다[2].

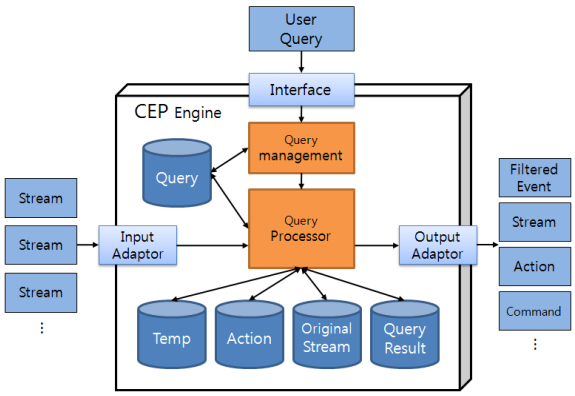
빅 데이터와 같은 스트림 데이터(Stream Data)를 처리하는 대표적인 솔루션으로 CEP(Complex Event Processing)가 있다. CEP 는 실시간으로 입력되는 스트림 데이터를 처리하는 미들웨어(Middleware)로, 정의된 질의에 해당되는 이벤트만을 선별하여 다양한 패턴 등의 규칙을 탐사할 수 있다[3][5]. 그러나 CEP 는 정의되지 않은 잠재적인 중요 이벤트에 대한 규칙 탐사에는 한계가 있다.

이 논문에서는 CEP 환경에서 빅 데이터 사이에 존재하는 규칙 탐사를 위한 실시간 데이터 마이닝 기법의 개발을 제안한다. 이 논문의 구성은 다음과 같다. 2 장에서는 CEP 의 구조와 연구현황에 대하여 살펴보고, 3 장에서는 CEP 의 기능적 제약사항을 기술한다. 4 장에서는 제안하는 CEP 환경 기반의 실시간 데이터 마이닝 기법을 기술하고, 5 장에서 결론 및 향후 연구를 기술한다.

## 2. CEP 의 구조와 연구현황

다양한 종류의 센서 네트워크를 통하여 수집되는 빅 데이터는 하나의 이벤트로 정의된다. 그리고 다양한 종류의 센서와 영역으로부터 실시간으로 동시에 수집되는 이벤트는 매우 방대하고 복잡하다. CEP 는 실시간으로 수집된 데이터를 처리하기 위한 미들웨어로 효율적인 빅 데이터 처리가 가능하고 처리 비용도 절감할 수 있다[4].

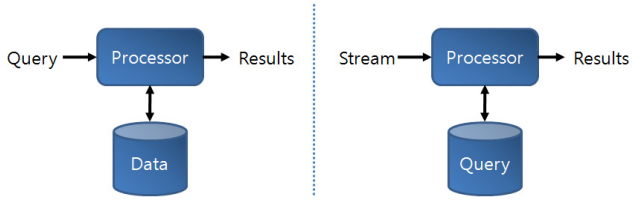
CEP 는 (그림 1)과 같이 다양한 종류의 스트림 데이터를 실시간으로 처리하여 다양한 종류의 서비스를 제공한다. CEP 는 실시간으로 수집된 복수의 이벤트들 사이에 존재하는 특정 패턴을 탐사하고 탐사된 패턴이 사용자가 관심을 갖는 패턴인 경우 사용자가 미리 등록한 질의(Query)를 수행한다[3].




(그림 1) CEP 의 구조

CEP 는 스탠포드 대학의 STREAM(The Stanford Data Stream Management System) 프로젝트에서 연구되었다[3]. STREAM 은 데이터 스트림 매니지먼트 시스템(DSMS)의 프로토타입(Prototype)을 정의하는 프로젝트로 전통적인 데이터베이스시스템에서 연속적으로 발생된 스트림 데이터 처리를 위한 연속 질의를 수행하여 데이터를 처리한다. 특히, 실시간으로 입력되는 스트림 데이터와 제한된 시스템 리소

스 환경을 고려하여 개발되었다. 또한 데이터 스트림을 처리할 수 있는 CQL(Continuous Query Language)을 개발하였다. CQL 은 (그림 2)처럼 CEP 에 탑재되어 데이터를 처리하는 언어로, 데이터베이스의 데이터 처리언어인 SQL 구문과 유사한 방식으로 정의된다[3][5][6].




(그림 2) SQL 과 CQL 의 비교

## 3. CEP 의 기능적 제약사항

CEP 의 핵심 기능은 다음과 같다.

첫째, CEP 는 실시간으로 입력되는 스트림 데이터를 사용자가 사전에 정의한 질의를 수행하여 이벤트와 패턴을 탐사한다[3]. 그러나 실시간으로 다양하게 생성되는 이벤트들 사이에 존재하는 정의되지 않은 의미가 있는 잠재적인 이벤트 및 패턴에 대해서는 능동적으로 처리하지 못하는 한계가 있다. 즉 CEP 는 데이터 처리에 있어서 정의된 질의에 대한 이벤트와 패턴만을 탐사하는 수동적인 방식이다.

둘째, CEP 는 패턴 매칭(pattern matching) 기술을 제공한다[3]. 현재 제공되는 패턴 매칭에는 정제(Filtering), 집합(Aggregation), 상관관계(Correlation), 선행관계(Causality) 등이 있다. 패턴 매칭을 통해 사용자가 정의할 수 있는 이벤트 패턴에는 동시 발생(and), 발생하지 않음(not), 이벤트 발생 후(A followed-by B) 등과 같은 패턴이 전부이다. 따라서 기존의 CEP 구조는 특정 패턴의 빈발함(support), 신뢰도(confidence)와 같은 조건 적용에 제한이 있다.

셋째, CEP 는 데이터 처리 언어인 CQL 을 제공한다[3]. 사용자는 CQL 로 CEP 에 미리 질의를 등록함으로써, 관심이 있는 이벤트와 패턴에 대한 정보를 얻을 수 있다. 현재 사용되는 다양한 종류의 솔루션에서 적용하는 CQL 은 단순 패턴 매칭 기능만을 제공하고 있다. 따라서 기존에 정의된 CQL 로는 다양한 패턴 매칭을 표현하는 데 한계가 있다.

## 4. CEP 환경 기반의 실시간 데이터마이닝 기법

현재의 CEP 는 실시간으로 수집되는 빅 데이터들 사이 존재하는 새로운 규칙 탐사에 한계가 있다. 이러한 제약사항은 기업의 긴급한 비즈니스 서비스, 신속함이 중요시 되는 다양한 의사결정 과정에서 큰 문제가 될 수 있다. 따라서 사용자가 사전에 등록해 놓은 질의에 대한 이벤트나 패턴뿐만 아니라, 시간의 흐름에 따라 새롭게 발생된 중요한 이벤트와 패턴을 탐사할 수 있는 데이터 마이닝 기법을 적용한 지능적인 CEP 엔진이 개발되어야 한다.

기존의 CEP 에 데이터 마이닝 기법을 적용하기 위한 방법으로 다음과 같은 과정을 수행이 필요하다. 먼저 고급 패턴 매칭을 수행할 수 있는 시스템 모델을 개발해야 한다. 기존의 CEP 에서 제공하는 패턴 매칭은 "event A 의 값이 10 이상인 경우 질의를 수행하라", "event A 가 발생한 다음 event B 가 발생한 경우 질의를 수행하라"와 같은 매우 단순한 기본 기능만을 제공한다. 따라서 "event A 와 event B 가 동시에 발생한 경우가 전체 이벤트의 30% 이상인 경우 질의를 수행하라"와 같은 고급 패턴 매칭 모델을 개발할 필요가 있다. 이러한 고급의 패턴 매칭 모델을 개발한다면, 새로이 발생하는 규칙과 패턴을 탐지할 수 있다. CEP 스스로 다양한 규칙과 패턴을 탐지하게 된다면, 지금보다 더 다양한 응용 분야에서 적용될 수 있다.

다음으로 CEP 를 위한 실시간 데이터 마이닝 기법을 개발해야 한다. 고급 패턴 매칭 기능을 구현하더라도, 곧바로 적용하는 것은 불가능하다. 패턴 매칭 기능을 CEP 에 적용하려면, 고급 패턴 매칭 기능을 위한 실시간 데이터 마이닝 기법을 새로이 개발해야 한다.

마지막으로 기존의 CQL 을 확장한다. 기존의 CQL 은 고급 패턴 매칭 모델과 실시간 데이터 마이닝 기법을 표현하는데 한계가 있다. 예를 들어, Sybase 사에서 개발한 Aleri SQL 의 경우, 선행관계(Causality)를 표현하기 위해 "A fby B"와 같이 " fby" 키워드를 사용한다. EsperTech 사의 Esper 의 경우 "A -> B"와 같이 "->"으로 선행관계를 표현한다. 이처럼 기존의 CQL 은 패턴 매칭의 기법들을 완벽하게 표현하지 못하고 있다. 따라서 개발한 기법들을 CEP 엔진에서 실행하려면, 기존의 CQL 을 확장하는 연구가 필요하다.

## 5. 결론 및 향후 연구

이 논문에서는, CEP 의 개념과 연구 현황을 살펴보고, CEP 의 기능과 한계점을 기술하였다. CEP 에 데이터 마이닝 기법을 적용하면, 새롭게 정의되는 이벤트와 패턴 탐사에 능동적인 대처가 가능할 것이라 예상한다.

향후 연구로 구체적인 데이터 마이닝 기법을 접목시킨 지능형 CEP 를 개발하고자 한다. 그리고 성능 분석을 통하여 기존의 CEP 와의 성능을 확인하고 새롭게 정의되는 이벤트와 패턴 탐사에 대한 실시간 규칙 탐사 능력을 분석하고자 한다.

## 참고문헌

[1] Manyika, j., Chui, M.,Brown, B., Bughin, J., dobbs, R., Roxburgh, C., and Byers, A. H. "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company. 2011.

[2] 이만재. "빅 데이터와 공공 데이터 활용", Internet and Information Security, 제 2 권, 제 3 호, pp.47~64, 2011.

[3] Arasu, A. and Babcock, B. and Babu, S. and Cieslewicz, J. and Datar, M. and Ito, K. and Motwani, R and Srivastava, U. and Widom, J. "*STREAM: The Stanford Data Stream Management System*". Stanford InfoLab. 2004

[4] 강만모, 구자록, 이동형. "차세대 웹 환경에서 Complex Event Processing 엔진을 이용한 대용량데이터 처리", 정보과학회논문지: 데이터베이스 제 37 권 제 6 호. 2010.

[5] S. R. Madden, M. A. Shah, J. M. Hellerstein, and V. Raman. "Continuously Adaptive Continuous Queries over Streams", Proc. of ACM SIGMOD 2002, Madison, Wisconsin, United States, 2002.

[6] http://www.cs.brown.edu/research/aurora