KIPS Tr. Comp. and Comm. Sys.
Vol.3, No.10 pp.377~382 pISSN: 2287-5891

http://dx.doi.org/10.3745/KTCCS.2014.3.10.377

The Method for Extracting Meaningful Patterns Over the Time of Multi Blocks Stream Data

Kyeong-Rae Cho† · Ki-young Kim††

ABSTRACT

Analysis techniques of the data over time from the mobile environment and IoT, is mainly used for extracting patterns from the collected data, to find meaningful information. However, analytical methods existing, is based to be analyzed in a state where the data collection is complete, to reflect changes in time series data associated with the passage of time is difficult. In this paper, we introduce a method for analyzing multi-block streaming data(AM-MBSD: Analysis Method for Multi-Block Stream Data) for the analysis of the data stream with multiple properties, such as variability of pattern and large capacitive and continuity of data. The multi-block streaming data, define a plurality of blocks of data to be continuously generated, each block, by using the analysis method of the proposed method of analysis to extract meaningful patterns. The patterns that are extracted, generation time, frequency, were collected and consideration of such errors. Through analysis experiments using time series data.

Keywords : Multi-Block Streaming, Internet of Things, Big Data Analysis, Continuance Data Analysis, Time Series Data, Digital Native

시간의 흐름과 위치 변화에 따른 멀티 블록 스트림 데이터의 의미 있는 패턴 추출 방법

조 경 래† · 김 기 영††

요 약

모바일 통신과 사물 인터넷(IoT) 환경에서 시간에 따른 데이터의 분석 기술은 주로 의미 있는 정보를 찾기 위해 수집 된 데이터에서 의미 있는 패턴을 추출하기 위해 사용된다. 기존의 데이터 마이닝을 이용한 분석 방법은 데이터 수집이 어렵고 시간의 경과와 관련된 시계열 데이터의 변경을 반영하기 위해 완료 상태에 기초하여 해석되어야 한다. 이러한 패턴의 다양성, 대용량성, 연속성 등의 여러 가지 특성을 가진 데이터 스트림의 분석을 위한 방법으로 멀티 블록 스트리밍 데이터 분석(AM-MBSD) 방법을 제안한다. 의미 있는 데이터 추출을 위해 멀티 블록 스트리밍 데이터의 패턴을 추출하고 추출된 연속적 데이터를 여러 개의 블록으로 정의하고 제안 방법의 검증을 위해 각 데이터 블록의 데이터 패턴 생성 시간, 주파수를 수집하고 시계열 데이터를 분석, 실험하였다.

키워드 : 멀티 블록 스트림, 사물 인터넷, 빅데이터 분석, 연속성 데이터 분석, 시계열 데이터, 디지털네이티브

1. 서 론

최근 스마트폰 도입이 일상화되면서 이른바 모바일 컴퓨팅이나 지능형 시스템 등과 같은 지능적인 서비스를 필요로 하는 워크플레이스의 근본적인 변화가 생활 전반에서 일어나고 있으며 특히 스마트 모바일 환경에서 다양한 데이터 분석 기법이 유용하게 사용되고 있다. 최근 데이터 마이닝(Data Minning) 기법은 멀티미디어들의 다양한 특성에 맞게 세분화 되고 각 도메인(Domain) 별 수집된 기존 데이터로부터 유용한 정보를 추출하기 위해서 사용되는데, 데이터를 수집하고 수집된 방대한 양의 다양한 데이터 패턴을 추출[1]하여, 그 정보를 기반으로 데이터 분석요구에 따른 적절한 서비스를 제공하는 것이 주요 이슈로 부각되고 있다. 하지만, 지금까지의 데이터 마이닝 기법은 수집이 완료된 상태에서 다양한 데이터를 분석하는 것을 기반으로 하고 있다. 더욱이 수집된 데이터 안에서 시간의 흐름과 위치의 변화에 따라 연속적으로 변화하는 패턴이 존재하는 경우, 이를 반영한 최적의 분석 기법은 찾아보기 어렵다. 예를 들어, 사람

† 정 회 원 : 서일대학교 컴퓨터소프트웨어과 조교수
†† 정 회 원 : 서일대학교 컴퓨터소프트웨어과 부교수
Manuscript Received : September 12, 2014
Accepted : October 8, 2014
* Corresponding Author : Kyeong-Rae Cho(krcho@seoil.ac.kr)

이 태어나서 어린 나이에 디지털 기술을 이용하는 연령과 컴퓨터 사용이 익숙하고 스마트폰 통신사의 인터넷을 사용할 수 있는 디지털네이티브(Digital Native, 디지털 융합 기기 사용자)들이 생겨나면서 사물의 인터넷 생태계(ECO)가 조성되고 있다. 이들의 송, 수신 정보를 이용한 통화 패턴을 분석한다고 가정할 때, 단위 시간당 발생하는 통화정보의 양은 매우 방대한 빅데이터(Big Data)일 것이다. 또는 CCTV와 같은 영상 감시 장치로부터 실시간으로 재난 발생 가능성과 이상 징후를 사전에 판단한다고 할 때, 짧은 시간에 축적되는 여러 종류의 대용량의 데이터에서 패턴의 유사도 분석을 통해 유용한 패턴을 추출하는 것은 어려운 일이다. 뿐만 아니라, 실시간 대기 오염도 측정 및 원자력 발전소의 원자로 온도 모니터링과 같은 지능형 유/무선 센서 네트워크 매체 환경에서부터 유/무선 네트워크 환경에서 패킷 분석을 통한 스마트 모바일 콘텐츠 융합기술 서비스 분야[4]에 이르기까지 매우 다양하다. 또한 이러한 연속성을 가지는 데이터의 특징 중에는 시간의 흐름과 디지털네이티브(디지털 융합 기기 사용자)의 위치 변화에 따라 스마트폰 이용 패턴의 변화가 발생할 수 있는데, 이와 같은 패킷 전송의 연속성, 대용량성 그리고 스마트폰 사용자의 사용 패턴의 변화 등의 특성을 가진 데이터를 스트림 데이터(Stream Data)라고 하며, 이러한 데이터를 데이터의 종류에 따라서 각 데이터는 도메인별로 분류하여 멀티 블록 스트림 데이터를 만들어 전송하며, 여러 종류의 데이터 객체에 대한 멀티 블록 데이터 스트림의 특징을 다음과 같이 분류할 수 있다.

첫째, 대용량성이다. 이는 멀티 블록 스트리밍 데이터 형태로 전송되는 스트리밍 데이터의 크기(Size of Data)가 너무 방대하기 때문에 분석이 복잡하고 분석한다고 하더라도 분석 후 추출된 결과를 해석하기 어렵다.

둘째, 데이터의 연속성(Streaming Data)이다. 스마트폰을 사용하는 디지털네이티브의 요구에 따라 필요한 데이터 수집이 완료된 상태가 아니라 시간(위치)의 변화에 따라 연속적인 데이터가 계속해서 축적되고 있다.

셋째, 시간의 흐름에 따른 데이터 패턴의 가변성(Variable of Pattern)으로 분류할 수 있다. 또한 실시간 분석의 어려움을 해결하기 위하여 단순하게 저장 공간에 축적하는 것은 분석 할 수 없을 정도의 많은 양의 데이터를 만들게 된다. 또한 수집된 멀티 블록 스트리밍 데이터는 시간의 흐름과 위치의 변화에 따라 실시간적으로 변화하는 패턴을 가질 수 있으며, 패턴의 변화에 적응적인 서비스가 요구된다. 예를 들어, 지능형 스마트 모바일 콘텐츠 사용자에게 성향에 맞게 멀티 블록 데이터를 선택적으로 맞춤형 콘텐츠를 제공한다고 가정하자. 이때 먼저 선행되어야 하는 것은 시간 변화에 따른 디지털네이티브의 위치 및 장소를 파악하는 과정이 요구된다. 이를 디지털네이티브의 시간변화 및 위치변화에 따른 데이터분석 과정이라고 말하며, 데이터분석 과정의 결과물을 디지털네이티브가 일상생활에서 스트림 데이터를 사용하는데 있어서 무슨 데이터를 어떻게 소비하는지 소비방식을 찾기 위한 유용한 데이터 분석 추출 모델이라고 한다. 하지만, 유용한 데이터 추출 모델은 시간과 디지털네이티브의 위치에 따라 변화한다. 스트리밍 데이터의 반복적인 패턴 분석 과정에서 유용한 데이터 추출 정확도를 높이게 된다. 이에 따른 디지털네이티브의 콘텐츠 이용 성향도 변할 수 있기 때문에, 본 논문에서는 데이터의 대용량성, 연속성, 패턴의 가변성의 특성을 가지는 멀티 블록 스트림 데이터의 분석을 위해 유용한 패턴 추출을 위한 멀티 블록 스트림 데이터 분석 방법(AM-MBSD: Analysis Method for Multi-Block Stream Data)을 제안한다. AM-MBSD는 여러 종류의 멀티 블록 스트리밍 데이터를 분석하기 위해 데이터를 임계 구역(Multi-$Block_{th}$)으로 구분하고, 제안 분석 방법을 통해서 Multi-$Block_{th}$을 분석하여 유용한 패턴을 추출한다.

멀티 블록 구간에서 Multi-$Block_{th}$에 분석된 결과는 데이터의 분석 파라미터로 데이터 발생시간, 빈발(Frequent), 데이터 모호성(불확실성) 등을 고려하여 조합하였다.

본 논문의 구성은 다음과 같다. 2장에서는 스트림 데이터의 유용한 패턴 추출을 위한 분석과 관련된 연구에 대해서 소개하고, 3장에서는 데이터의 순서나 시간의 흐름에 따라 스트림 데이터의 대용량성, 연속성 요소, 시계열 정보와 같은 패턴의 변화 요소에 따라 다른 특징을 가지는 멀티 블록 스트림 패턴의 변화의 다른 특징을 나타내는 멀티 블록 스트림 데이터의 형태를 정의한다. 4장에서는 유용한 패턴 추출을 위한 멀티 블록 스트리밍 데이터 분석 방법(AM-MBSD: Analysis Method for Multi-Block Streaming Data)을 설명하고 멀티 블록 데이터를 비교, 분석실험을 통해 제안 기법이 적절한 결과를 도출해내는 것을 확인한다. 마지막으로, 5장에서는 결론 및 향후 연구과제에 대해서 논한다.

2. 멀티 블록 스트림 데이터의 패턴 변화 형태

본 논문은 멀티 블록 스트림 데이터의 대용량성, 연속성 및 패턴의 변화 요소는 대용량성이나 연속성 요소와 다른 특징을 가진다. 멀티 블록 스트림 데이터는 대용량성과 연속성 또는 시계열 정보를 가지는 데이터에서 순서나 시간의 흐름에 따라 패턴의 변화를 가지는 형태라고 정의할 수 있다. 이런 데이터는 도메인(Domain)에 따라서 다양하게 확인할 수 있는데, 예를 들어, 실생활에서 스마트 모바일 융합 콘텐츠를 이용하는 디지털네이티브들의 콘텐츠 응용분야나 관심의 변화와 다수 디지털네이티브의 성향 변화, 지식 습득 수준 등을 높일 수 있다. 또는 모바일 게임 환경에서 게임 이용자의 사용자 경험(UX: User Experience) 또는 사용자 인터페이스(UI: User Interface)에 대한 활용 친숙도에 따른 운용(Operation) 능력 등을 말한다. 이와 같은 패턴 분류는 다음과 같이 여러 가지 응용 도메인의 응용 형태에 따라 데이터 추출 후 형태를 분류하면 표준형(Standard Type), 규칙적 변화형(Regular Variations), 불규칙(또는 비예측) 변화형(Irregular Variations), 규칙적 반복형, 불규칙적 반복형으로 구분할 수 있다.

Table 1. Data extraction and analysis form classification

구 분	설 명
표준형	패턴의 변화가 없는 일관된 형태
규칙적 변화형	패턴 데이터가 규칙적으로 변화하는 형태
불규칙적 변화형	패턴이 불규칙적으로 변화하는 형태
규칙적 반복형	과거 데이터가 규칙적으로 반복되는 형태
불규칙적 반복형	과거 패턴이 없이(랜덤하게) 불규칙적으로 반복되는 형태

표 1은 표준형(Standard Type)이 일관된 형태로 패턴이 포함되어 있어서 불특정 구간에 발생한 데이터를 이용하여 분석해도 동일한 결과가 추출된다. 그렇기 때문에 일반적으로 데이터 분석에 많이 사용되는 형태이다. 즉, 표준형은 이진(Binary) 스트리밍 데이터를 이용하여 "과거 정보기반으로 $(000/001/000/001)_2$ 즉, $(0101)_{10}$을 선택했다"는 패턴을 추출했다면 "미래에 추출될 결과도 역시 과거 패턴과 동일하게 $(000/001/000/001)_2$ 즉, $(0101)_{10}$을 선택할 것이다"라고 추이 또는 추정하는 것이다. 하지만, 규칙적 변화형(Regular Variations)은 "과거 정보를 기반으로 $(000/001/010/011)_2$ 즉, $(0123)_{10}$을 선택했다"는 패턴을 추출했다면 "미래에는 $(100/101/110/111)_2$ 즉, $(4567)_{10}$을 추출할 것이다"라고 추이 또는 추정하는 형태이다. 이처럼, 규칙적 변화형은 분석 방법과 목적이 기존 표준형과 차이를 가진다.

규칙적 반복형은 패턴 변화가 없는 일관된 형태로 과거의 패턴을 주기적으로 나타내고, 규칙적 반복형(Regular Iterative)은 과거 데이터가 규칙적으로 반복되는 형태이다. 불규칙적 변화형(Irregular Variations)과 불규칙적 반복형(Irregular Iterative)은 과거 패턴 데이터가 주기적으로 불규칙하게 자주 변화하고, 시간의 흐름에 따라 비예측적으로 반복되는 형태로서 광의의 의미로는 표준형과 유사한 분석 방법이긴 하지만, 데이터가 규칙적이지 않기 때문에 패턴을 추출하기 힘든 형태라고 할 수 있다[10].

본 논문에서는 규칙적으로 변화하고 반복되는 데이터의 형태에 대해서 구체적으로 논의 하겠다. 그림 1a, 그림 1b, 그림 2a, 그림 2b는 표 1에서 분류한 도메인 활용 형태에 따른 패턴 데이터의 변화를 비트 스트리밍의 형태로 표현하였다.

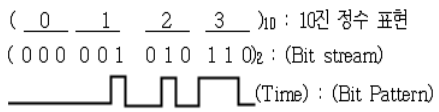


Fig. 1a. Standard form (historical data pattern)

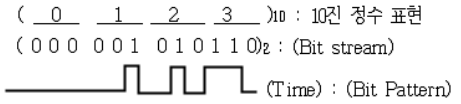


Fig. 1b. Standard form
(Historical patterns and future trends are the same pattern)

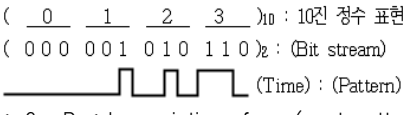


Fig. 2a. Regular variations form (past pattern)

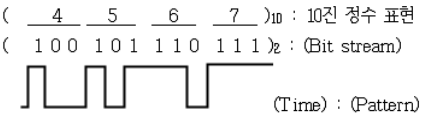


Fig. 2b. Regular variations form
(Representing the difference between the pattern and the past pattern, the future trend)

3. 유용한 패턴 추출을 위한 멀티 블록 스트리밍 데이터 분석 기법(AM-MBSD: Analysis Method for Multi-Block Streaming Data)

3.1 유용한 패턴 추출 기법

본 논문에서 제안하는 유용한 패턴 추출을 위한 멀티 블록 스트리밍 데이터 분석 방법(AM-MBSD: Analysis Method for Multi-Block Streaming Data)은 대용량성과 연속성의 특성을 가지는 과거데이터에서 패턴의 변화를 고려하여 분석하는 기법이다. AM-MBSD를 이용하기 위해서는 먼저 데이터의 양과 발생 시간 및 간격 등을 고려하여 멀티 블록 임계 영역(Multi-Block_{th})의 범위를 결정해야 한다. 수집되는 멀티 블록 스트리밍 데이터가 Multi-Block_{th}의 크기를 만족할 때, 과거에서 현재까지 시간의 흐름에 따라 발생하는 데이터와 그 데이터를 Multi-Block_{th} 영역으로 결합하여 규칙을 만들고 만들어진 규칙을 결합한다. 이런 과정을 통해서 추출된 유용한 패턴 데이터를 응용하고자 하는 분야, 즉 도메인(Smart mobile Contents, Social Network Service/System, m-Learning, Game, mobile Electric Vote, Multimedia Selection, etc)을 이해하고 활용하려는 요구에 따라 해결과제에 따라 속성을 결정한다. 결정된 속성 및 파라미터에 따라서 데이터 분석에 사용하기 위해 필요한 멀티미디어 데이터(Text, Audio, Video(image), Animation, Game, etc.)를 수집한다. AM-MBSD는 디지털네이티브가 소비하는 여러 종류의 데이터 블록 스트림에 대한 소비(이용) 패턴의 유사성을 분석하여 패턴 추출 결과를 재조립한 뒤 응용 도메인별로 분류하여 재조립하는 기법이다.

각 서비스(IPTV, VOD, VCS, m-Learning 등)를 분류하기 위해서 다음과 같이 네 가지 단계를 통해서 각 응용 분야(도메인)에 필요한 데이터를 분석하게 된다.

첫째, 데이터 수집과 유용한 패턴 추출을 위한 멀티 블록 스트리밍 데이터 임계 구역들은 여러 종류의 멀티 블록 스트리밍 데이터를 분석하기 위해 데이터를 멀티 블록 임계 영역(Multi-Block_{th})으로 구분하여 저장하고, Multi-Block_{th}의 크기는 시계열 데이터나 데이터의 종류와 데이터의 크기로 지정할 수 있으며, 디지털네이티브가 응용하려는 분야(도메인)의 특성에 따라 상대적으로 분류된 블록 스트림 데이터를 분류한다.

둘째, 제안 모델인 AM-MBSD을 이용해서 패턴 간 유사성을 찾기 위해서는 멀티 블록 스트리밍 데이터 분석을 위해 데이터의 양, 발생시간 및 발생 간격과 같은 속성 값들을 정의하고 멀티 블록마다 임계 영역으로 모아진 데이터를 이용하여 패턴을 추출하여 패턴 간의 유사성을 찾고, 추출된 패턴을 모아서 유사한 패턴끼리 같은 도메인에 있음을 결정하는 과정이다. AM-MBSD에서는 분석 결과를 if-then 형식의 블록 구조로 변환하여 저장하고 관리한다.

셋째, 멀티 블록 스트리밍 데이터 규칙의 결합(Rule Set)은 각 블록(임계 영역, Multi-Block$_{th}$) 단위와 구간마다 발생한 멀티 블록별 패턴 규칙들을 패턴 간의 유사성 분석을 통해서 얻어진 패턴 규칙을 통합하여 조립(Assemble) 후 다시 동일 응용 분야 도메인별로 분류하는 과정을 의미한다. 규칙을 결합할 때는 규칙 간의 오류나 포함 관계를 고려할 수 있어야 하고, 각각의 멀티 블록의 스트리밍 데이터 규칙의 결합 과정에서 발생하는 빈발(Frequent) 횟수, 단일 블록 구

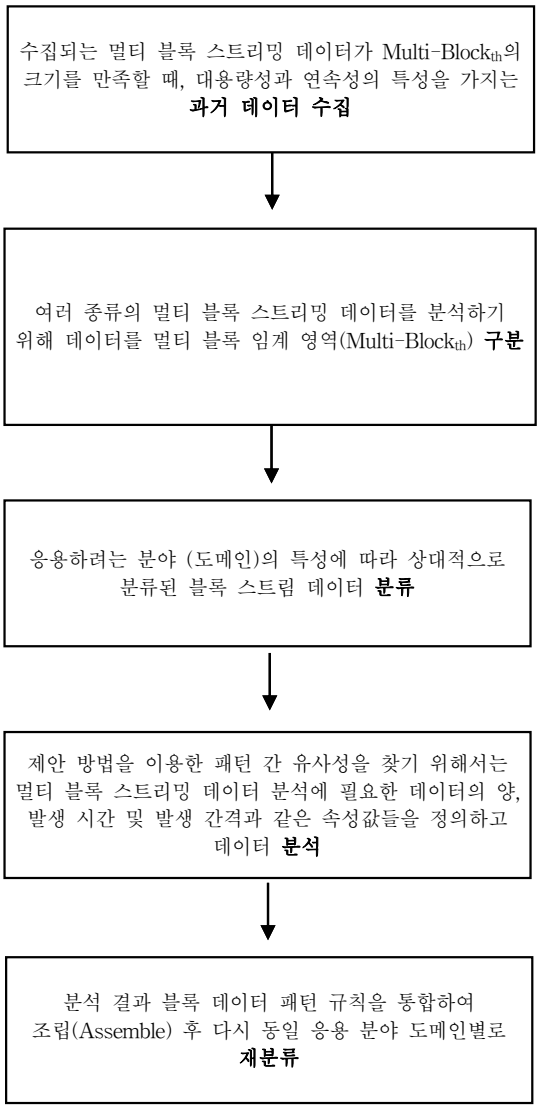


Fig. 3. A flow chart for extracting a useful pattern Multi-block stream

간에서 사용된 최대 반복시간, 그리고 에러 빈발(Frequent) 횟수를 이용하여 멀티 블록 스트리밍 데이터 규칙의 신뢰성을 측정한다. 또한 패턴의 변화를 반영하기 위해 규칙의 발생 시기를 반영할 수 있어야 한다. 그림 3은 유용한 패턴 추출을 위한 멀티 블록 스트리밍 데이터 분석 흐름도를 나타내고 있다[8]. 멀티 블록 스트림 데이터 규칙과 규칙 간의 유사형태에 따라서 합치형과 포괄형, 오류형으로 분류할 수 있다. 합치형의 경우 비교하고자 하는 두 규칙 간에 합치하거나 유사성이 높아서 둘 중에 하나의 규칙을 선택하여 사용하는 경우이다. 포괄형은 하나의 규칙이 또 다른 하나의 규칙에 일부 또는 모두가 포함되는 경우를 말한다.

하나의 규칙에는 여러 개의 조건을 포함하게 된다. 각 규칙사이에 포함되어 있는 조건들이 전체적으로 한 쪽 규칙에 포함될 수 도 있고, 부분적으로 포함관계를 보이는 규칙도 있다. 오류형의 경우 두 개의 규칙에 포함되어 있는 조건들은 동일하지만 결과가 서로 상이하게 나타남으로 모순이 발생한 경우이다. 이런 경우에는 규칙이 가지고 있는 신뢰도를 이용하여 높은 값을 가지는 규칙을 유지시킨다. 데이터 규칙의 신뢰도 검증은 멀티 블록 데이터를 비교 분석 실험을 통해서 확인하였다. 데이터 발생 빈발(Frequency of Occurrence, FO) 횟수, 최근 연속구간 반복시간(Time of Continuance, TC), 오류 빈발(Frequent of Noise, FN) 횟수를 이용하여 계산되며 유효성을 정량화(Quantification)하였다.

넷째, 패턴 데이터와 모델 생성은 축적하여 발생된 규칙의 집합(rule set)에 대하여 신뢰도와 유효성 수치를 기반으로 최종 결과물을 도출함으로써 유용한 패턴을 생성을 거쳐서 각 도메인에 필요한 응용서비스를 할 수 있도록 분석 하였다. 규칙의 신뢰도는 전체 규칙에 적용되는 블록 스트림 데이터 사용 빈발 횟수, 최근 연속 반복시간(시계열 데이터) 그리고 오류(Noise) 등을 반영하여 규칙의 집합에 가중치(Weights$_{rule_set}$)를 이용한 패턴을 분석한다. 데이터 규칙의 가중치 부여를 위한 알고리즘을 다음과 같이 표현할 수 있다.

$$Weights_{rule\,set} = \frac{FO}{FO_{Max}} + \frac{TC}{TC_{Max}} - FN$$

* FO: 발생빈발수($Frequency\ of\ Occurrence$)
* TC : 최근지속시간($Time\ of\ Continuance$)
* FN: 오류빈도수($Frequency\ of\ Noise$)

여기서 FO는 데이터 패턴의 발생 빈발수를 나타내며, 임계 영역(Multi-Block$_{th}$, MB)에서 얼마나 자주 빈발하게 발생했는지를 파악하기 위한 파라미터값으로 신뢰도가 높을수록 수치가 높다. TC는 최근 지속 시간으로 높은 가중치를 부여하면 최근 발생 빈도수가 높게 나타난다.

3.2 멀티 블록 데이터 비교 분석

실험의 목적은 방대한 양의 멀티 블록 스트림 데이터로부터 유용한 패턴(Pattern)과 규칙(Rule)을 추출해내기 위해

탐색(Exploration), 분석(Analysis)을 행한다. 데이터 마이닝(Data Mining)의 검증(Verification)과 발견(Discovery) 기법은 데이터로부터 새롭고 유용한 패턴을 추출하고 추출된 데이터 규칙의 신뢰도를 증명하기 위해 일부 객체들의 작용을 예측 가능하도록 데이터로부터 관련 패턴을 찾는다. 모든 데이터규칙에 대해서, 활용 빈발(Frequent) 횟수, 최근의 반복 지속 시간 그리고 오류에 반영하여 규칙에 가중치($Weights_{rule\_set}$)를 부여한다.

검증을 위해 데이터 발생 빈발(Frequent) 횟수는 각각의 임계 영역별 분석을 통해서 생성된 규칙들에서 얼마나 자주 동일한 규칙이 발생했는지 여부를 확인할 수 있다. 최근 지속시간은 과거에 발생했던 규칙 중에서 오류가 발생했던 경우에 대하여 패널티를 부여하기 위한 것이다. 규칙의 가중치 부여를 위한 규칙조건을 정의한다. 데이터 규칙의 신뢰도는 모든 데이터규칙에 대해서 최근의 데이터 반복 지속시간은 과거에 발생했던 데이터 규칙을 유용하게 사용하기 위한 것이다.

Table 2. Compare the number of data occurs frequently in the critical section with the rule of the Bit stream

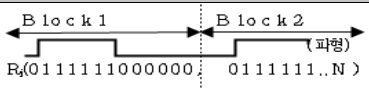
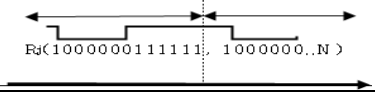
CASE	임계 영역 내에서 'N'까지 규칙 발생한 빈발(Ri, R_j)수
1	
2	

표 2는 임계 영역 내에서 N까지 데이터 규칙의 발생 빈발수는 동일하지만, 규칙이 발생된 시간이 경과함에 따라 규칙 빈발수가 다르게 나타나는 경우(발생하면, '1', 발생하지 않으면, '0')이다. 첫 번째 경우에는 R_i라는 데이터 비트(1)/비트 스트림이 처음 한 번 발생하지 않다가('0') 6번 동안 반복하여 발생('1')하더니 최근 6회 동안은 발생하였다. 두 번째 경우에는 R_j라는 비트 패턴이 처음 한 번 발생('1')하더니 6회 동안은 발생이 없었고, 이후 6회 동안 발생하는 패턴이 첫 블록에서와 같이 동일하게 반복하여 최근 다음 블록2에서도 규칙이 첫 번째 멀티 블록과 같이 반복하여 발생하였다. 이때 두 경우에 발생 빈발 수만 비교 한다면, 동일한 결과를 가져오지만, 최근에 발생한 두 번째 멀티 블록에 우선순위가 높은 가중치($Weights_{rule\_set}$)를 부여한다면, 당연히 첫 번째 경우가 규칙이 될 것이다. 실험을 통해서, 두 가지 경우의 멀티 블록 데이터 비트패턴은 시간의 흐름과 각 비트 사이의 위치(구간)의 변경에 따라 반복적이고 규칙적인 변화를 가지는 멀티 블록 데이터를 비교 분석하였으며, 기존 데이터 마이닝을 이용한 방법보다 멀티 블록 스트림 단위로 분석하여 실험한 경우, 데이터 규칙(rule_set)의 개수가 감소하면서 데이터 패턴 구별이 용이한 패턴을 추출할 수 있었고, 평균 에러율이 낮아진 모습을 확인할 수 있었다.

4. 결 론

스마트 모바일 융합기술의 발달과 함께 여러 환경에서 수집되는 데이터의 크기나 형태 및 속성들이 다양한 모습을 보이고 있다.

제안 방법인 AM-MBSD은 사물 인터넷 기반의 다양한 스마트 모바일 콘텐츠 기술 응용 분야에서 수집되는 과거 데이터를 모두 이용하여 연속적이고 반복적인 패턴의 변화를 가지는 멀티 블록 형태의 비트 스트림 데이터를 더욱더 유용하고 효율적으로 분석 할 수 있는 기법을 소개하였다. 또한 멀티 블록 임계 영역($Multi\text{-}Block_{th}$)에 따라 디지털 네이티브의 개별적 요구에 따라 발생한 비트데이터 패턴은 연속적이면서 규칙성을 갖는 멀티 블록 스트리밍 데이터 분석에서 AM-MBSD 모델이 사물간통신 상황에서 발생되는 빅데이터를 예측하여 분석할 수 있는 방법으로 효과적임을 확인하였다.

향후 연구로는 임계 영역을 확장하여 추출된 멀티 블록 스트림 데이터 패턴을 분류하고, 분류된 패턴 규칙을 결합함으로써 좀 더 유용한 패턴의 신뢰성을 향상시킬 수 있는 보다 효율적인 기법을 제시함으로써 본 논문에서 제안하는 기법의 실제 응용 분야(도메인)를 확장하여 적용할 수 있는 실험을 할 것이다.

References

[1] Yon-sik Lee, Hyun Ko, "Kwangjong Kim, Extraction of Optimal Moving Pattern using Maximum Frequent 2-Sequence", Korean Institute of Information Scientists and Engineers, Vol.35, No.1(D), pp.367-362, 2008.

[2] Jaeeun Jung, "For mining of sensor network data streams Ontology-based pre-processing techniques", *Korea Intelligent Information System Society*, Vol.15, No.3, pp.67-80, 2009.

[3] Pponomartskoo, Youngjin Nam, Daewha Seo, "Homomorphic Cryptoschemes based Secure Data Aggregation for Wireless Sensor Networks", *Korean Institute of Information Scientists and Engineers*, Vol.35, No.1(A), pp.235-236, 2008.

[4] Hee-Sung Kim, Je-Young Lee, Jae Doo Park, Kwang Ho Choi, Young Joon Lee, Jong-Joon Choi, Min Woo Kim, Hyung Keun Lee, "A Conceptual Study to Share Real-Time Data Stream from Continuously Operating GPS/GNSS Reference Stations", *Korean Institute of Information Scientists and Engineers*, Vol.35, No.1(A), pp.136-142, 2009.

[5] P. Domingos and G. Hulten, "Mining high-speed data streams", In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery ND Data Mining, 2000.

[6] "Data mining techniques", http://www.aistudy.co.kr/ learning/ mining/technique_jang.htm#_bookmark_19d3d80

[7] Yong-yoon Suh, Hak-yeon Lee, Yon-tae Park, "Analysis and Visualization of Structure of Smartphone Application Service Using Text-Minimg and the set-covering algorithm", *International journal of Mobile Communications*, Vol.10, No.1, pp.1-20, 2012.

[8] Yong-Jun Cho, Joon Hur, "A Study on Improving the Predict Accuracy Rate of Hybrid Model Technique Using Error Pattern Modeling: Using Logistic Regression and Discriminant Analysis", *Journal of the Korean Data & Information Science Society*, Vol.17, No.2, pp.268-278, 2006.

[9] Hong-Cheol Lee, "A Study on the join algorithm of neural networks (C4.5) decision tree for the customers classification of mobile communication", *Journal of the Korea Intelligent Information System Society*, Vol.9, No.1, pp.139-155, 2003.

[10] Tae-Bok Yoon, Jee-Hyong Lee, Kyeong-Rae Cho, "Advanced Learner's Modeling based on Weighted Support Vector Data Description(SVDD)", *Journal of the Korean Institute of Information Scientists and Engineers*, Vol.40, No.1, pp.46-52, 2013.

[11] J. Xi, "Outlier Detection Algorithms in Data Mining", *IEEE Second Int. Symp. on Intelligent IT Application*, Vol.1, pp.94-97, 2008.

조 경 래

e-mail : krcho@seoil.ac.kr

1999년 철도청 철도전산정보사무소 정보화 연구원

2000년 성균관대학교 정보통신공학과 (연구과정)

2003년 성균관대학교 정보통신공학과 (공학석사)

2009년 성균관대학교 컴퓨터공학과(공학박사)

2009년~2012년 THEMOST 한국IT융합기술연구소장

2012년~현 재 서일대학교 컴퓨터소프트웨어과 조교수

관심분야 : 인간과 컴퓨터 상호작용(HCI), ICT융합공학, 디지털 콘텐츠, 클라우드 및 빅데이터 분석, 사물 인터넷, 지능형 센서 네트워크, 스마트 모바일 컴퓨팅, 컴퓨팅 보안과 개인정보보호, 인터넷윤리 등

김 기 영

e-mail : ganet89@seoil.ac.kr

1996년 상지대학교 전자계산학과(이학사)

1995년~1997년 삼보정보통신 기술 연구소 연구원

1999년 숭실대학교 컴퓨터학과(공학석사)

2003년 숭실대학교 컴퓨터학과(공학박사)

2004년~현 재 서일대학교 컴퓨터소프트웨어과 부교수

관심분야 : 모바일 컴퓨팅, 센서네트워크, 네트워크보안, 사물 인터넷