계절성이 갖는 시계열 데이터의 특성 탐지 방법에 대한 연구

오재훈* · 김두진* · 김종달*

서 론

웹 클릭스트림 데이터의 폭발적 증가로 인하여, 분석해야 할 데이터 양이 증가하고 있으며 데이터 조합의 경우의 수는 기하급수적으로 증가하고 있다. 따라서, 시계열 데이터 분석의 경우에도 필요한 분석 조합의 수가 수동적으로 분석할 수 있는 한계를 넘어서고 있어, 시계열 데이터의 특성을 자동적으로 분석할 수 있는 시스템이 요구되고 있다.

그림1 의 그래프를 보면 유사한 패턴이 주기적으로 반복되고 있다는 사실을 파악할 수 있다. 이와 같은 변화가 생기는 원인은 인간의 행동이 계절에 따라 변화하기 때문이다. 이러한 이유로 온라인에서 발생하는 데이터들 중에서도 계절성을 가지는 데이터가 많이 발생한다.

계절성이란 기후변화, 공휴일의 위치, 기업관례, 전망에 의해 발생하는 것으로 불변의 규칙적인 형태는 아니지만 스펙트럼의 계절적 주기 부근에 정점이 나타나게 하는 1년 이내의 체계적인 움직임을 말한다([이한식])

그림1에서 1부터 211까지는 계절적 특성이 매우 유사하고, 최대값이 커지는 경향을 보인다. 하지만, 211 이후에는 최대값이 작아지는 반면에 폭이 커지는 것을 알 수 있다.

이와 같이, 계절성을 갖는 시리즈 데이터를 그래프로 표시하면 우리는 즉각적으로 데이터의 계절성을 파악할 수 있다. 분석가는 그래프의 오목한 부분과 볼록한 부분을 인식하고 이 패턴들이 반복되는지를 분석한다.

계절성 시계열 데이터 분석과 관련된 문제가 두 가지가 있다. 첫째는, 데이터의 증가로 인해 분석해야할 시계열 데이터의 수가 기하 급수적으로 증가하고 있다는 것이다. 따라서, 시계열 데이터의 모든 조합을 수작업으로 일일이 확인하는 것이 불가능해지고 있다.

둘째로는, 분석가들이 이 데이터의 계절성의 특징에 대한 질문을 받는다면 정확히 대답하기 어려워 한다는 것이다. 대부분의 분석가들은 주기적 반복 패턴과 패턴의 변화에 대해서 체계적으로 분석하기 보다는 임의적이고 임기응변식의 방식으로 해석하는 경우가 많기 때문이다.

이와 같은 문제를 해결하기 위해서는 다음과 같은 질문에 체계적으로 답할 수 있는 방법이 필요하다.

* ㈜넷스루 데이터마이닝 연구소

1. 이 데이터가 계절성을 가지고 있는가?
2. 계절성의 반복 주기는 얼마나 되는가?
3. 반복되는 패턴의 간격이 줄어들고 있는가? 늘어나고 있는가?
4. 전체적인 측면에서 그래프가 상승추세인가? 하향 추세인가? 아니면 정체상태인가?
6. 반복되는 패턴의 높이는 커지고 있는가? 작아지고 있는가?
7. 반복되는 패턴의 면적은 증가하고 있는가? 감소하고 있는가?

2절에서는 시계열 데이터 분석과 관련된 기존 연구들에 대해서 살펴본다. 3절에서는 시계열 데이터의 특성을 정의하고, 특성을 저장하고 관리하기 위한 데이터 모델에 대해서 설명한다. 4절에서는 시계열 데이터의 특성을 분석하기 위한 알고리즘을 설명한다. 마지막 5절에서는 본 논문의 의미를 살펴보고 향후 연구과제를 제시한다.

2. 관련 연구

시리즈 데이터의 특성에 대한 연구는 신호 처리와 데이터 마이닝 분야에서 많이 다루어졌다. 특히, 데이터 마이닝에 대한 연구가 활발했던 1990년대 이후 시계열 데이터의 특성에 대한 연구가 활발히 이루어졌다.
데이터마이닝과 빅 데이터와 관련된 연구들에서는 시계열 데이터를 저장하고, 시계열 데이터의 특성을 이용하여 시계열 데이터를 조회하기 위한 기술들이 연구되어왔다([Stam]). 하지만, 이러한 연구의 대부분은 시계열 데이터에서 의미가 적은 데이터를 제거하여 데이터를 압축하는 방법과 이를 저장하고 조회하는 방법에 대한 연구가 주를 이루었다([Motro], [Shasha] [Agrawal1]). 데이터 조회시에도 유사도를 이용한 방법이 주로 연구되었다.
시리즈 데이터의 계절성에 대해서도 많은 연구들이 이루어졌으나, 계절성을 가진 데이터를 체계적으로 해석하고 분석하고 이를 저장하고 조회하는 방법에 대한 연구는 거의이루어지지 않았다.

3. 시계열 데이터의 모델

본 섹션에서는 시계열 데이터의 특성을 조회하기 위한 데이터 모델을 설명한다. 시계열 데이터 모델은 시계열 원본 데이터를 저장하기 위한 테이블, 시계열에 대한 요약정보를 저장하기 위한 테이블, 시계열의 계절적인 반복주기에 대한 정보를 저장하기 위한 테이블로 구성된다.

3.1 시계열 테이블

시계열 데이터를 저장하는 테이블명은 SERIES 이다. 시계열 데이터 테이블에는 시계열 데이터의 요약정보를 저장하고 관리하며, 테이블을 구성하는 컬럼은 “표. 시계열 데이터 테이블”과 같다.

컬럼	설명
SERIES_ID	시계열 데이터를 식별하기 위한 식별자
NAME	시계열 데이터의 이름
START	시계열 데이터가 시작되는 지점
END	시계열 데이터가 끝나는 지점
COUNT_VALUE	시계열 데이터의 수

MAX_VALUE	시계열 데이터의 최대값
MIN_MVALUE	시계열 데이터의 최소값
SUM_VALUE	시계열 데이터들의 합
AVG_VALUE	시계열 데이터들의 평균
STD_DEV	시계열 데이터들의 표준편차
INTERVAL_COUNT	시계열 반복구간 수

표 1. 시계열 데이터 정보 테이블 (SERIES)

3.2 시계열 반복구간 테이블

시계열 데이터에서 반복되는 구간을 저장하는 테이블을 시계열 반복구간 테이블(SERIES_INTERVAL)이라 한다. 시계열 데이터를 구성하는 반복구간에 대한 정보를 저장하고 관리하며, 테이블을 구성하는 컬럼은 “표. 시계열 반복구간 테이블” 과 같다.

컬럼	설명
INTERVAL_ID	시계열 구간 식별자
SERIES_ID	시계열 구간이 속하는 시계열 테이블의 식별자
START	구간이 시작되는 지점
END	구간이 끝나는 지점
LENGTH	구간의 길이. 구간이 시작되는 지점과 구간이 끝나는 지점 사이의 간격
MAX_VALUE	구간에서 시계열 데이터의 최대값
MIN_MVALUE	구간에서 시계열 데이터의 최소값
SUM_VALUE	구간의 시계열 데이터들의 합
AVG_VALUE	구간의 시계열 데이터들의 평균
STD_DEV	구간의 시계열 데이터들의 표준편차
PEAK_POINT	구간에서 가장 최대인 지점의 위치

표 2. 시계열 데이터 반복구간 테이블

3.3 시계열 데이터

시계열 데이터 테이블은 시계열 원본 데이터를 저장하기 위한 정보를 저장하고 관리하며, 테이블을 구성하는 컬럼은 “표. 시계열 데이터”와 같다.

컬럼	설명
SERIES_ID	시계열 데이터를 식별하기 위한 식별자
SERIES_INDEX	시계열 데이터의 인덱스
SERIES_VALUE	시계열 데이터의 해당 지점 값

표 3. 시계열 데이터

4. 시리즈 데이터의 계절성 분석

시계열 데이터의 변동성분을 다음과 같이 네가지로 분류한다.

- 추세변동 : 변동이 증가하거나 감소하는 장기적인 경향 (예: 인구, 물가 변동)
- 순환변동 : 상승과 하락이 주기적으로 중기적으로 나타나는 변동으로 계절변동에 비해 주기성이 적으나 지속적으로 나타나는 정기변동 (예: 불황, 호황)

- 계절변동 : 기후조건, 사람의 습성, 휴일 등의 계절적 요인, 기타 주기적인 요인으로 인해 반복적으로 나타나는 변동
- 불규칙변동 : 예측할 수 없는 임의의 변동

본 절에서는 추세변동이나 순환변동이 거의 없는 계절변동을 가진 시계열 데이터의 특징을 추출하고, 이를 저장하고 조회하는 방법을 제안한다.

4.1 시계열 데이터 특징(feature)

시계열 데이터에서 의미를 찾고 이를 저장하고 조회하기 위해서는 먼저 시계열 데이터의 특징을 정의하여야 한다. 본 논문에서는 계절성의 변화를 다음과 같은 특성으로 정의한다.

1. 계절성 반복 구간(Interval) : 계절변동이 반복되는 간격
2. 폭 : 반복 구간의 길이를 나타냄
3. 높이: 반복 구간 안에서 최대값
4. 면적 : 반복구간의 구간의 합
5. 표준편차 : 반복구간 데이터의 표준편차

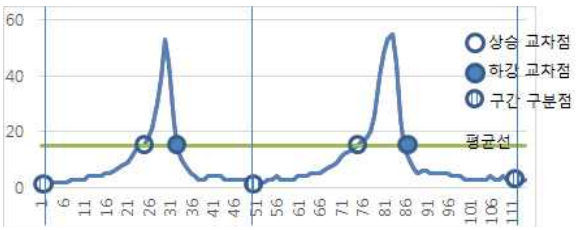

그림 2. 계절성 반복 구간

계절성을 갖는 시계열 데이터에서 위와 같은 정보를 추출하려면 계절성 반복구간을 계산하기 위한 알고리즘이 필요하다. 본 논문에서는 평균선을 이용하여 계절적 특징들을 추출하는 방법을 제시한다. 반복구간을 계산하기 위해 필요한 용어를 다음과 같이 정의한다.

- 평균선 : 시계열 전체 데이터의 평균값
- 교차점(cross point) : 평균선과 시계열 데이터 곡선이 교차하는 지점. 교차점은 상승교차점과 하강 교차점 두 종류가 있다.
- 상승교차점(increasing cross point) : 교차점에서 시계열 데이터가 우상승하는 경우
- 하강교차점(decreasing cross point) : 교차점에서 시계열 데이터가 우하강하는 경우
- 구간 구분점 : 계절성이 반복 구간을 나누는 기준점. 이전 반복 구간이 끝나는 지점인 동시에, 새로운 반복 구간이 시작되는 지점이다.

4.2 반복 구간 탐지 알고리즘

반복구간의 특징을 추출하려면, 시계열 데이터에서 반복구간을 식별하여야 한다. 반복구간을 식별하기 위한 알고리즘은 다음과 같다.

시계열 데이터 S = { $x_1, x_2, ..., x_n$ } 에 대해서

1. 시계열 데이터 전체의 평균을 구한다. (FindAverage)
2. 시계열 데이터 곡선과 평균선이 만나는 교차점을 구한다. (FindCrossPoint)
3. 교차점을 스캔한다. 교차점이 상승교차점이면, 이전 교차점과 현재 교차점 사이의 최소값을 가지는 지점을 구한다. 이 지점이 구간을 구분하는 지점이 된다. (FindIntervalPoint)
4. 각 구간에 대해서 구간의 특징에 해당되는 값들을 계산한다. (FindIntervalFeatures)

FindAverage : 전체 평균 구하기

$$average = \frac{\sum_{i=1}^{n} x_i}{n}$$

FindCrossPoint: 원본 시계열의 평균값과 시계열 데이터 곡선이 교차하는 교차점 목록을 구한다.

- k : 현재까지 찾은 교차점을 나타내는 색인값
- average : 원본 시계열의 평균값
- x_i : 원본 시계열에서 i 번째 값
- cp_k : k 번째 교차점(cross point)
- cp_k.cross point type : k 번째 교차점의 종류 (상승교차점, 하강교차점)
- cp_k.index : k 번째 교차점이 원본 시계열상의 위치

```
k = 0
for = 0 to n-1 {
  if ( x_i < average and average < x_{i+1} ) {
      cp_k.cross point type = 상승교차점
      cp_k.index = i;
      k++;
  }
  if ( x_i > average and x_{i+1} < average ) {
      cp_k.cross point type = 하강교차점
      cp_k.index = i;
      k++;
  }
}
```

FindIntervalPoint : 교차점 목록을 스캔하면서 구간 구분점을 구한다.

- j : 현재까지 찾은 구간 구분점을 나타내는 색인값
- x_i : 원본 시계열에서 i 번째 값
- pp_j : i 번째 구간 구분점(interval point)
- pp_j.index : j 번째 구간 구분점이 원본 시계열 데이터 상의 위치

```
j = 0
// x_1 을 구간 구분점 목록에 추가한다.
pp_0.index = 1;

// for every cross point
for ( k=0; k < cp.size ; k++ ) {
  if ( cp_k 가 하강 교차점 ) {
      i=cp_i.index
      intervalMinValue = x_i
      intervalMinIndex = i;
      for ( ; i<cp_{i+1}.index; i++ ) {
          if ( x_i < intervalMinValue ) {
              intervalMinValue = x_i
              intervalMinIndex = i;
          }
      }
      pp_j.index = intervalMinIndex;
      j++;
  }
}
```

FindIntervalFeatures : 구간 구분점들을 스캔하면서 구간에 대한 정보를 구한다.

- pp.size : 구간 구분점 목록의 크기
- pp_i.index : i 번째 구간 구분점이 원본 시계열 데이터 상의 위치
- period : 구간 특징 정보를 저장하기 위한 변수

for (i=0; i < pp.size ; i++) {
  s=pp_i.index ;
  e=pp_{i+1}.index ;
  interval.max_value = max($x_s, x_{s+1}, \ldots, x_e$)
  interval.min_value = min($x_s, x_{s+1}, \ldots, x_e$)
  interval.sum_value = sum($x_s, x_{s+1}, \ldots, x_e$)
  interval.period_count = e - s;
  interval.avg_count = sum_value / (e-s);
  interval.std_dev = stddev($x_s, x_{s+1}, \ldots, x_e$)

5. 시계열 데이터 조회

계절성을 가진 시계열 데이터베이스가 구축이 되면, 이 데이터베이스를 이용하여 다양한 형태의 질의를 사용할 수 있다. 계절성에 대한 데이터 접근 패턴은 다음과 같이 분류할 수 있다.

- 시계열 데이터 내부의 특징 질의 : 시계열 데이터 내부의 특징을 조회하기 위한 질의
- 시계열 데이터간 비교 질의 : 서로 다른 시계열 데이터 사이의 특징을 비교하기 위한 질의

본 절에서는 시계열 데이터 내부의 특징을 질의하는 방법만을 설명한다. “표. 시계열 데이터의 구간 테이블“은 네이버 트렌드에서 텐트로 검색했을 때의 검색 데이터를 분석하여 얻은 반복구간 테이블이다. ”텐트“라는 검색어 데이터베이스를 이용하여 다음과 같은 질의를 던질 수 있다.

1. 텐트가 최대치를 기록하는 계절(월)은 언제인가?

```
SELECT A.NAME, B.PEAK_POINT
FROM SERIES A, SERIES_INTERVAL B
WHERE A.SERIES_ID = B.SERIES_ID
```

2. 텐트의 계절성 편차를 내림차순으로 구한다.

```
SELECT A.NAME, B.STD_DEV
FROM SERIES A, SERIES_INTERVAL B
WHERE A.SERIES_NAME = “텐트”
AND A.SERIES_ID = B.SERIES_ID
ORDER BY B.STD_DEV DESC
```

3. 여름철에 최대치를 달성한 구간정보를 찾는다.

질의 :

날짜	구간1	구간2	구간3	구간4	구간5	구간6	구간7
구간 시작	20070101	20071231	20090216	20100111	20110124	20111226	20121224
구간 종료	20071224	20090209	20100104	20110117	20111219	20121217	20140127
길이	52	59	47	54	48	52	58
최소	2	2	2	4	4	4	4
최대	53	55	71	100	66	58	56
합계	451	614	713	912	833	1023	1118
평균	8.67	10.41	15.17	16.89	17.35	19.67	19.28
표준편차	11.04	12.87	17.96	21.16	16.24	16.65	15.09
피크지점	20070723	20080728	20090727	20100719	20110718	20120723	20130610

표 4. “텐트” 검색어 시계열 데이터의 구간 테이블

```
SELECT A.NAME, B.*
FROM SERIES A, SERIES_INTERVAL B
WHERE A.SERIES_NAME = "텐트"
AND A.SERIES_ID = B.SERIES_ID
 AND 7 <= B.PEAK_POINT.MONTH <= 8
```

6. 결론 및 향후 연구 과제

본 연구에서는 계절성을 갖는 시계열 데이터의 특징을 분석하는 알고리즘을 제시하였다. 계절성을 가지고 있는 경우에는 반복주기 탐색, 폭의 변화, 높이의 변화, 면적의 변화, 표준편차와 같이 반복 구간의 특징을 데이터베이스로 구축하고, 이를 조회하기 위한 방법도 제안하였다.

이 시스템을 이용하면 분석가들이 수동으로 분석하지 않아도 시계열 데이터의 계절적 특성을 분석하고 이를 데이터베이스로 구축할 수 있다. 또한, 계절성 특징을 이용해 시계열 데이터를 조회할 수 있기 때문에 분석가들은 단순 시계열 분석이 아닌 다양한 형태의 질의를 수행할 수 있다.

본 연구는 추세변동이나 순환변동이 거의 없고 계절변동만을 가진 시계열 데이터를 대상으로 진행되었기 때문에, 계절변동이 분명한 경우에는 매우 잘 작동하지만, 추세변동이나 순환변동 혹은 불규칙 변동이 많은 경우에는 잘 작동하지 않는다. 따라서, 시계열 데이터의 특성을 저장하고 조회하는 시스템의 완성도를 높이기 위해서는, 추세변동이나 순환변동이 있을 경우에 모델을 확장하기 위한 연구가 수행되어야 한다.

참 고 문 헌

[1] 이한식, 경제 시계열자료의 계절성 분석: 계절모형 접근방법의 개관, 계량경제 학보 제11권 제3호 117-157, 2000

[2] R. Stam and R. Snodgrass, "A Bibliography on Temporal Databases", IEEE Bulletin on Data Engineering, 11(4), Dec. 1988.

[3] A. Motro, "VAGUE: A User Interface to Relational Databases that Permits Vague Quries," ACM Trans. on Information System (TOIS), 6(3), pages 187-214, July 1988.

[4] D. Shasha and T-L. Wang, "New techniques for best-match retrieval", ACM TOIS, 8(2):140-158, April 1990.

[5] R. Agrawal, T. Imielinski, and A. Swami, "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning and Discovery in Knowledge-Based Databases

[6] E. Keogh, S.Chu, D. Har, M. Pazzani, Segmenting Time Series: A Survey and Novel Approach

[Fu] Tak-Chung Fu, A Review on Time Series Data Mining, Engineering Applications of Artificial Intelligence, 2011

[7] C.S Perng, H.Wang, S.R.Zhang, D.S.Parker, Time Series Change Detection using Segmentation: A Case Study for Land Cover Monitoring

K. B. Pratt, Locating Patterns in Discrete Time Series, 2001

[8] R.Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search In Sequence Databases", Research Report, IBM Almaden Research Center, San Jose, Californica, 1993

[9], R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami, "An Interval Classifier for Database Mining Applications", VLDB 92, Vancouver, August 1993

오 재 훈

• 1995년 포항공과대학교 컴퓨터공학과 (공학석사)
• 1995년-2000년 포항공과 자동화센터 전임연구원
• 2000년-2008년 ㈜넷스루 개발팀장
• 2008년-현재 ㈜넷스루 데이터마이닝 연구소장
• 관심분야: 클릭스트림 분석, 빅 데이터 분석, 데이터 마이닝, 개인화, 추천

김 종 달

• 2002년 포항공과대학교 컴퓨터공학과(공학석사)
• 2000년-현재 ㈜넷스루 데이터마이닝 연구소
• 관심분야: 빅데이터 분석, 데이터 마이닝, 클릭 스트림 데이터 분석

김 종 달

• 1995년 포항공과대학교 컴퓨터공학과(공학석사)
• 1994년-2000년 포스데이타
• 2000년-현재 ㈜넷스루 데이터마이닝 연구소
• 관심분야: 빅데이터 분석, 데이터 마이닝, 클릭 스트림 데이터 분석, 개인화 추천