동적으로 변화하는 정보에 대한 모니터링 및 적응적 변화 예측

박대욱, 이원석
연세대학교 컴퓨터과학과
e-mail : dwpark@database.yonsei.ac.kr
leewo@database.yonsei.ac.kr

Monitoring and adaptive prediction of the dynamically changed information

Dae Wook Park, Won Suk Lee
Dept. of Computer Science, Yonsei University

요 약

최근의 온라인 응용 환경에서는 다양한 종류의 데이터 스트림을 다루고 있으며 이러한 데이터 스트림은 빠른 속도로 무한히 생성되고 실시간의 빠른 처리를 필요로 한다. 따라서 데이터 스트림 실시간 처리 및 분석 작업에서는 데이터 스트림을 지속적으로 모니터링하여 앞으로의 변화와 이에 따른 부하를 예측하고 성능을 조절하는 일이 필요하다. 본 논문에서는 끊임없이 발생하는 데이터를 관찰하여 데이터가 발생하는 패턴을 찾아내고, 찾아낸 패턴을 기반으로 미래의 특정 시점에서 발생할 데이터 값을 미리 예측하는 효율적인 기법을 제안한다. 무한한 양의 데이터를 제한된 크기의 메모리 내에서 처리하여 현재부터 과거 특정시점까지 발생한 데이터의 패턴을 가장 정확히 일반화할 수 있는 함수를 찾아내고 그 함수를 기반으로 미래에 발생할 데이터의 값을 예측한다.

1. 서론

최근 센서 네트워크나 e-비즈니스 및 주식 시장의 온라인 분석 등의 응용 환경에서는 다양한 종류의 데이터 스트림을 다루고 있다. 이러한 데이터 스트림은 빠른 속도로 끊임없이 무한히 생성되지만 실시간의 빠른 처리를 필요로 한다. 따라서 데이터 스트림 실시간 처리 및 분석 작업에서는 데이터 스트림을 지속적으로 모니터링하여 앞으로의 변화를 예측하고 이로 인해 발생할 수 있는 부하를 예측하고 성능을 조절하는 과정이 요구된다.

본 논문의 목적은 끊임없이 발생하는 데이터를 모니터링하여 데이터가 발생하는 패턴을 찾아내고, 찾아낸 패턴을 기반으로 미래의 특정 시점에서 발생할 가능성이 있는 데이터 값을 미리 예측하고자 하는 것이다. 입력 데이터는 시간의 경과에 따라 무한히 발생하기 때문에 그 양이 방대하다. 그러나 이러한 데이터를 다루기 위한 시스템의 메모리 크기는 한정되어 있기 때문에 본 논문에서는 각 주기의 실제 데이터를 모두 저장하지 않고 해당 주기의 패턴을 주기함수화하여 일반화된 주기함수의 매개 변수만을 저장한다. 일반적으로 데이터의 값은 한정된 범위 내에서 발생한다고 가정할 수 있으므로, 현재부터 과거 특정 시점까지 발생한 데이터의 패턴을 가장 정확히 일반화할 수 있는 함수를 찾아내어 현재 모니터링 되고 있는 데이터의 특성을 파악할 수 있다. 또한 그 함수를 기반으로 미래에 발생할 데이터의 값을 예측할 수 있다.

* 이 논문은 2007 년도 정부(과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업으로 수행된 연구임 (No.R0A-2006-000-10225-0).

2. 관련 연구

데이터 스트림에서의 변화 탐지를 정의하기 위해서는 먼저 '변화'가 무엇인지를 명확히 할 필요가 있으며 이와 관련된 연구가 진행되어 왔다[1,2]. 직관적으로, 데이터 스트림의 발생 과정이 변화하면 데이터 스트림으로부터 유발되는 특징들 또한 반드시 변하게 된다. 따라서 데이터 스트림의 특징에 대한 변화가 탐지되면 데이터 스트림의 발생 과정이 변화되었음을 알 수 있다. 또한 여러 실제 응용 환경에서, 데이터 스트림 발생 메커니즘의 변화는 항상 스트리밍 데이터 분포의 변화를 의미한다. 일반적으로 데이터는 특정 분포를 따라 발생된다고 가정할 수 있기 때문에, 변화 탐지는 데이터 분포의 변화를 탐지하는 문제로 좁혀지며 이와 관련되어 여러 연구들이 이루어졌다[1,3,4]. 분포의 차이를 탐지하는 알고리즘 중 가장 널리 사용되는 알고리즘은 Wilcoxom test[5], Lp distance와 Jensen-Shannon Divergence[6] 등이 있다. 한편 변화 탐지 및 예측에 관한 방법들도 제안되었다[1,7,8]. [1]은 오차 추정에 기반한 변화 탐지 방법을 제안하였으며, [8]은 탐지된 변화에 대한 통계적 신뢰도를 보증하

나 고차원 데이터 스트림에 대해서는 비실용적이다.

3. 변화하는 정보에 대한 모니터링 및 변화 예측

3.1 주기 함수 기반의 데이터 발생 패턴 탐색

데이터 스트림 환경에서 발생되는 데이터는 시간 x에 대한 함수 $f(x)$로 볼 수 있다. 연속적인 시간에 대하여 $f(x)$ 값이 일정한 범위 내에서 발생할 때, 그 발생 형태에 따라 여러 가지 함수로 일반화시킬 수 있다. 특히 발생하는 데이터가 임의의 주기마다 증가와 감소를 반복한다고 가정할 때, 이러한 패턴을 가장 잘 표현할 수 있는 함수는 주기 함수라고 할 수 있다. 따라서 본 논문에서는 데이터 발생 패턴을 주기 함수의 형태로 일반화하고자 한다.

데이터의 발생 패턴을 *sine*, *cosine*, *tangent* 등과 같은 주기 함수로 일반화하기 위해서는 입력 데이터로부터 각 함수가 필요로 하는 매개 변수를 구해야 한다. 예를 들어, *sine* 함수는 다음 식과 같이 일반화할 수 있으며, 세 개의 매개 변수 α, β, γ가 필요하다.

$$f(x) = \alpha \cdot \sin \beta x + \gamma$$

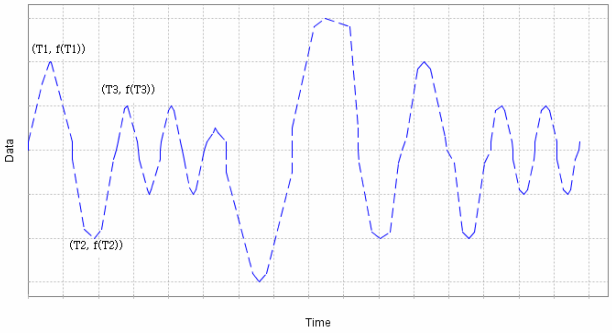
입력 데이터가 증가와 감소를 반복하면서 *sine* 함수와 비슷한 형태로 발생한다면, 미분계수가 양(+)에서 음(-)으로 바뀌는 순간을 기준으로 주기를 구분할 수 있을 것이다. 이러한 기준에 따라 각 구간을 구분하고 난 후 해당 구간에 대하여 세 개의 매개 변수를 구하게 되면 입력 데이터를 *sine* 함수로 일반화할 수 있게 된다. 그 외에도 여러 가지 다양한 기법으로 데이터 발생 패턴을 주기 함수의 형태로 일반화할 수 있다.

주기 함수의 하나인 *sine* 함수를 예로 들어 데이터 발생 패턴을 어떻게 일반화할 수 있는지 구체적으로 살펴보면, *sine* 함수의 일반형은 다음과 같다.

$$f(x) = \alpha \cdot \sin \beta x + \gamma \quad (\alpha > 0,\ \beta > 0)$$

위 식에서 α는 *sine* 함수의 진폭, β는 함수의 주기, 그리고 γ는 최대, 최소값의 중간을 의미한다. 즉 최대, 최소값은 $\gamma \pm \alpha$가 되며 주기는 $\frac{2\pi}{|\beta|}$가 된다. 그러나 데이터가 항상 일정한 *sine* 함수를 따라 발생한다고 보장할 수 없으므로 발생 패턴에 부합하는 매개 변수 값이 매 순간 달라질 수 있다. 따라서 매개 변수 α, β, γ가 정의되는 시점을 명시적으로 정의해야 할 필요가 있다.

sine 함수의 경우 주기마다 최대, 최소값이 반복하여 나타나므로 우리는 최대(최소)값이 발생하는 시점부터 최소(최대)값이 발생하는 시점까지, 즉 일반적인 *sine* 함수 주기의 $\frac{1}{2}$에 해당하는 구간을 *sine* 함수의 시작과 끝으로 볼 수 있다. 다시 말해, 함수의 시작과 끝은 데이터가 최대, 최소가 되는 시점으로 정의되며 이 시점은 좌표평면에서 그래프의 기울기(미분계수)가 0 이 되는 순간이다. (그림 1)에서처럼 $x = T1$ 에서 $f(T1)$이 발생하고 $x = T2$ 에서 $f(T2)$가 발생한다면 두 (x , y) 좌표 순서쌍 ($T1$, $f(T1)$), ($T2$, $f(T2)$)으로부터 $T1$, $T2$ 사이의 기울기를 구할 수 있다. 데이터의 기울기를 계속 모니터링 하면서 기울기가 0 으로 변하거나 양수에서 음수로, 또는 음수에서 양수로 바뀌는 시점 t_l를 찾아내고 같은 방식으로 기울기가 0 이 되는 다음 시점 t_m을 찾아내면 구간 $[t_l, t_m]$이 *sine* 함수로 일반화되는 구간이며, t_l과 t_m를 *sine* 함수의 최대, 최소로 간주하고 그 구간에 해당하는 α, β, γ를 구할 수 있다.



(그림 1) 시간의 흐름에 따른 데이터 발생

(그림 1)과 같이 데이터가 발생할 경우, $T1$에서 기울기가 최초로 0 이 되며 다음으로 0 이 되는 시점은 $T2$가 되므로 ($T1$, $T2$) 구간에 대한 매개 변수 α, β, γ를 다음과 같이 구할 수 있다.

$$\alpha = \frac{(f(T2) - f(T1))}{2}$$

$$\beta = \frac{2\pi}{(T2 - T1) \times 2}$$

$$\gamma = \frac{(f(T1) + f(T2))}{2}$$

위와 같은 방식으로 ($T2$, $T3$) 구간에 대해서도 α, β, γ를 구할 수 있으며 이 과정을 반복하여 모든 데이터 발생 구간에서 각각의 매개 변수를 구할 수 있다.

3.2 감쇄 기법의 적용

과거에 발생했던 데이터에 기반하여 미래의 데이터를 예측하기 위해 본 논문에서는 감쇄 기법(*Decay Mechanism*)[9]을 적용한다. 데이터 스트림 환경에서는

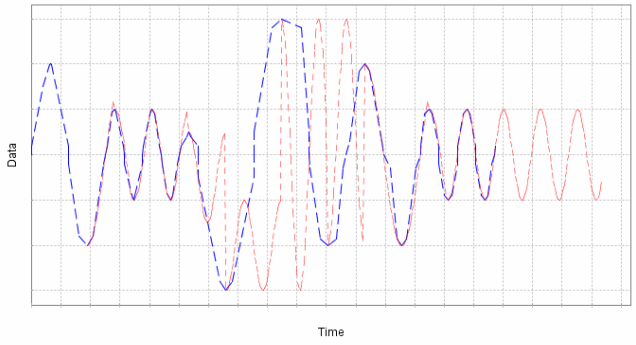
메모리 용량이 제한되어 있는 반면 데이터는 무한히 발생하므로, 과거의 모든 데이터를 영구적으로 유지하지 않고 가장 최근에 발생한 데이터에만 관심을 가지며 과거의 데이터는 오래된 것일수록 그 의미가 퇴색되었다고 본다. 따라서 이미 오래 전에 발생했던 데이터와 최근 발생한 데이터의 비중을 다르게 둘 수 있도록, 매개 변수마다 감쇄 기법을 적용하기로 한다.

감쇄값은 *decay-base b* 와 *decay-base-life h* 의 두 변수에 의해 정의된다. b 는 일정 시간이 지난 후 감쇄값이 줄어드는 비율을 정의한 것으로 정해진 시간이 지나면 감쇄값은 $1/b$ 의 비율로 줄어들게 되며 b 는 1 보다 큰 값을 가진다. h 는 감쇄값이 $1/b$ 로 줄어드는데 걸리는 시간을 의미하며, 가장 최근에 발생한 데이터 구간의 매개 변수의 감쇄값은 1 이 된다. 따라서 위의 두 변수에 의해 결정되는 감쇄율(*decay rate*) d 는 다음과 같이 정의할 수 있다.

$$d = b^{-(1/h)} \quad (b > 1, h \geq 1)$$

결과적으로 감쇄값은 0 에서 1 사이의 값을 가지며, b 와 h 를 변화시킴으로써 적용되는 감쇄율을 유연하게 조절할 수 있다.

어떤 경우에는 현재 발생한 매개 변수 값과 이전에 발생했던 매개 변수 값의 차이가 극도로 작을 수도 있다. 예를 들어 α_n =2.339 이고 α_{n+1} =2.338 이라면, α_n 과 α_{n+1} 의 차이는 1% 미만으로 매우 작다. 이런 경우 α_{n+1} 를 α_n 과 같은 값으로 봐도 무방하다. 그러므로 미리 사용자 정의값 δ 를 지정하여 어떤 매개 변수 값이 이전에 구한 어떤 매개 변수 값과 $\pm\delta$ (%) 범위 내에서 발생하게 되면, 두 값은 동일한 값으로 인식하도록 한다.



(그림 2) 데이터 발생 패턴 예측 결과

3.3 실시간 데이터 스트림 변화 예측

과거의 특정 시점부터 현재까지 발생한 데이터의 패턴은 세 개의 매개 변수 α, β, γ 로 함축되며 우리는 α, β, γ 를 이용하여 과거에 발생한 데이터의 패턴을 복원해 낼 수 있다. 감쇄값이 큰 매개 변수일수록 자주 발생했거나 가장 최근에 발생한 것이기 때문에 감쇄값이 가장 큰 매개 변수가 다음에도 발생할 확률이 가장 높다고 가정할 수 있으며 우리는 현재 갖고 있는 매개 변수들 중에서 감쇄값이 가장 큰 매개 변수를 선택하여 다음에 발생할 데이터를 예측할 수 있다. (그림 2)는 (그림 1)의 데이터를 *sine* 함수로 일반화하여 앞으로 발생할 데이터를 예측한 결과를 나타낸 그래프이다.

4. 결론

본 논문에서는 한정된 범위 내에서 발생하는 스트림 데이터에 대하여 과거 특정시점부터 현재까지 발생한 데이터의 패턴을 분석하여 주기함수의 형태로 일반화하고, 일반화된 함수를 기반으로 미래에 발생할 데이터의 값을 예측하는 알고리즘을 제안하였다. 또한 추출해 낸 주기함수의 매개 변수에 감쇄 기법을 적용하여 가장 최신의 데이터에 높은 가중치를 부여한다. 이와 같은 일련의 과정을 통해 앞으로 발생할 데이터의 패턴을 예측할 수 있다. 본 논문의 예제에서는 대표적 주기함수인 *sine* 함수만을 이용하여 데이터의 발생 패턴을 일반화하였지만 같은 방식으로 *sine* 함수뿐만 아니라 *cosine*, *tangent* 등 다른 여러 주기함수에 대해서도 일반화할 수 있다. 향후, 좀더 신뢰도 높은 변화 예측을 위해서 본 논문에서 언급되지 않은 오차에 대하여 수학적인 접근을 통한 연구가 이루어져야 할 것이다.

참고문헌

[1] Daniel Kifer, Shai Ben-David, Johannes Gehrke, Detecting Change in Data Streams, VLDB 2004: 180-191

[2] V. Ganti, J. Gehrke, R. Ramakrishnan, and W.-Y. Loh. A framework for measuring differences in data characteristics. Journal of Computer and System Sciences (JCSS), 64(3):542–578, 2002

[3] V.Ganti, J.Gehrke, and R.Ramakrishnan.Demon: Mining Data Streams under block evolution. SIGKDD Explorations,3(2):1-10,2002

[4] Sirsih Chandrasekaran, Michael Franklin: Remembrance of Streams Past: Overload-Sensitive Management of Archived Streams. In VLDB 2004

[5] F. Wilcoxon. Individual comparisons by ranking methods: Biometrics Bulletin, 1:80–83, 1945

[6] J. Lin: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151, 1991.

[7] Charu C. Aggarwal, A Framework for Change Diagnosis of Data Streams, SIGMOD Conference 2003: 575-586

[8] Wei Fan, Yi-an Huang, Haixun Wang, Philip S. Yu. Active Mining of Data Streams, SDM 2004

[9] J. H. Chang and W. S. Lee, Finding Recent Frequent Itemsets Adaptively over Online Data Streams, Proceedings of the 9th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, Washington, DC, August 2003 (ACM, New York, 2003) 487-492