# 패턴 변화 모델링과 모델 유사도에 관한 연구

A Study on Changed-Pattern Modeling and Model Similarity

저자
(Authors)

윤태복, 이지형, 최영미
TaeBok Yoon, Jee-Hyong Lee, YoungMee Choi

출처
(Source)

한국지능시스템학회 학술발표 논문집 20(2), 2010.11, 224-226(3 pages)

발행처
(Publisher)

한국지능시스템학회
Korean Institute of Intelligent Systems

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01546385

APA Style

윤태복, 이지형, 최영미 (2010). 패턴 변화 모델링과 모델 유사도에 관한 연구. 한국지능시스템학회 학술발표 논문집, 20(2), 224-226

# 패턴 변화 모델링과 모델 유사도에 관한 연구

# A Study on Changed-Pattern Modeling and Model Similarity

윤태복[1] · 이지형[2] · 최영미[3]
TaeBok Yoon, Jee-Hyong Lee, and YoungMee Choi

[12]성균관대학교 정보통신공학부
E-mail: [1] tbyoon@skku.edu, [2] jhlee@ece.skku.ac.kr
[3]성결대학교 멀티미디어공학부
E-mail: choiym@sungkyul.ac.kr

## 요 약

IT기술의 발달과 함께 환경에서 수집된 데이터는 대량성, 불완전성, 패턴의 변화 등의 특성을 가지고 있으며, 이런 데이터들의 특성을 고려한 고급화된 분석 기법이 요구되고 있다. 특히, 현실 세계 데이터의 패턴은 시간의 흐름에 따라 변화하는 모습을 빈번하게 보이고 있다. 본 논문에서는 내재된 패턴이 시간의 흐름에 따라 변화하는 데이터를 위한 분석 방법을 제안한다. 패턴 변화 구간을 엔트로피(entropy)를 이용하여 모델링하고, 시공간(Spatiotemporal)에 맵핑하여, 시간의 흐름에 따라 구간의 변화 패턴을 나타내었다. 인공 데이터를 이용하여 실험하여 적절한 결과를 확인하였다.

**키워드 :** Changed Pattern, Drifting Model, Model Similarity

## 1. 서 론

환경이나 사용자로부터 수집된 정보를 분석하여 의미 있는 패턴을 발견하고 미래를 예측하는 기술은 많은 분야에서 관심을 받고 있는 연구 분야이다. 학습자의 학습 성향 및 수준을 고려한 학습 시스템을 학습자에게 제공하기 위한 연구나 게임 환경에서 게임 플레이어의 게임 행위나 행동에 따른 게임 플레이어의 패턴을 인지하고, 그에 따른 적응된 게임 환경을 제공해 주기 위한 연구, 웹 검색 사용자를 위한 맞춤형/지능형 검색 결과 서비스, 스마트 폰 환경에서 사용자의 위치에 따른 적응형 추천 서비스 등이 대표적인 예라 할 수 있다. 하지만, 위와 같은 사례는 특정 도메인에 국한되지 않는 공통적인 전제를 가지고 있다. 과거에 발생했던 사례를 기반으로 하는 미래 예측이다. 바꾸어 말하면, 과거에 발생하지 않았던 상황이나 패턴에 대해서는 미래에도 반응할 수 없다. 이는 패턴이 변화하는 환경에서도 대응 할 수 없다는 의미로 확장된다. 앞의 예를 다시 이용한다면, 이러닝 환경이나 게임 환경에서 학습자나 게임 플레이어의 패턴이 항상 유지되지 않는 경우도 있다. 학습을 하면 할수록 학습자의 학습 수준은 향상 되고, UI 사용에 대한 패턴도 바뀔 수 있다. 게임 플레이어의 경우도 게임을 플레이 하면 할수록 능숙해지고, 반응도 빨라 질 것이다. 패턴의 변화를 인지하고 그에 맞는 미래를 예측하는 것은 어려운 작업일 것이다.

본 논문은 패턴 변화 정보를 하나의 모델로 생성하고, 패턴의 변화가 유사하게 발생하는 다른 모델로부터 미래를 예측 할 수 있는 정보를 얻는 것을 목적으로 한다. 패턴 변화를 위해 생성한 모델을 패턴 변화 모델 또는 Drifting model이라고 정의한다. 패턴 변화 모델 생성을 위해서는 임계구간을 이용하여 데이터를 분할하고, 엔트로피 값을 이용하여 n차원 공간에 사상(Mapping)시킨다. 공간에 사상된 모델간의 유사도를 측정하여 예측을 위한 배경 지식으로 활용한다. 제안하는 방법의 검증을 위한 실험에서는 2차원 기반의 다양한 형태를 가지는 인공데이터를 이용하였으며 적절한 결과를 확인하였다.

## 2.관련 연구

패턴의 변화를 가지는 환경에서 관련 연구는 다음과 같다. Aggarwal은 패턴 변화 데이터를 분석하기 위한 다양한 방법을 소개하고 있다. 그는 패턴 변화 데이터를 분석하기 위하여 군집(clustering), 분류(classification), 빈도 패턴 추출, 패턴 변화, 인덱싱 기법, 슬라이딩 기법 등 여러 가지 유용한 방법을 조사하였다[1]. Babcock 등은 패턴 변화 데이터 환경에서 모델 생성과 이슈에 대하여 소개하였고[2], Jain은 다양한 환경에서의 패턴 변화 데이터 발생을 소개하고, 통계적 방법을 이용한 패턴 변화 데이터 분석을 제안하였다[3]. 김진화와 민진영은 연속 발생 데이터의 실시간 분석을 위해 슬라이딩 윈도우(sliding window)를 이용하여 분할하고 의사결정나무 방법을 이용하여 분석하였다[4]. 하지만, 각각의 슬라이딩 윈도우에서 분석을 통하여 만들어진 규칙을 통합하는 과정이 명확하지 않고 단순하게 처리하여 효율성이 낮다. Golab 등은 패턴 변화 데이터의 저장/관리를 위한 다양한 방법을 제시하였다[5]. Domingos는 연속적인 데이터 발생 환경에서 효과적인 분석을 위한 Very fast decision tree(VFDT)를 제안하였다[6]. VFDT는 호페딩 트리 방법에 기반을 두고 있으며, 매우 빠른 데이터의 발생에 효과적으로 처리 할 수 있는 방법을 제안한다 하지만, 연속성 데이터의 분석을 고려할 뿐, 시간에 따른 데이터 패턴의 변화는 고려하기 어렵다는 문제를 가지고 있다.

## 3. 패턴 변화 모델링과 모델 유사도

### 3.1 패턴 변화 모델링

패턴 변화 모델은 수집된 데이터의 변화 흐름을 모델

링한 것이다. 수집 데이터(표 1)를 시간 또는 크기에 따라 분할하고(t), 각각 분할된 데이터의 ***knnWeight*** 고려하여 엔트로피를 계산한다. 각 분할 공간은 속성에 따라 엔트로피 값을 가지고 있으며, 이 값의 변화를 n차원 공간에 사상시킨다. ***n***차원 공간에 사상된 엔트로피 변화 정보를 패턴 변화 모델이라고 정의한다.

표 1. 수집 데이터 사례와 t에 따른 분할

| | $A_1$ | ... | $A_n$ | C | w | |
|---|---|---|---|---|---|---|
| $I_1$ | 20 | ... | 23 | Y | 0.8 | $t_1$ |
| $I_2$ | 12 | ... | 19 | Y | 0.3 | |
| $I_3$ | 25 | ... | 28 | N | 0.2 | |
| ... | ... | ... | ... | ... | ... | |
| $I_n$ | 23 | ... | 23 | Y | 0.4 | |
| $I_{n+1}$ | 2 | ... | 18 | Y | 0.8 | $t_2$ |
| $I_{n+2}$ | 25 | ... | 34 | N | 0.9 | |
| ... | ... | ... | ... | ... | ... | |
| $I_m$ | 64 | ... | 23 | Y | 1.5 | |
| ... | ... | ... | ... | ... | ... | |
| $I_{x+1}$ | 33 | ... | 34 | Y | 1.8 | $t_n$ |
| $I_{x+2}$ | 28 | ... | 23 | N | 0.8 | |
| ... | ... | ... | ... | ... | ... | |
| $I_z$ | 31 | ... | 21 | Y | 0.3 | |

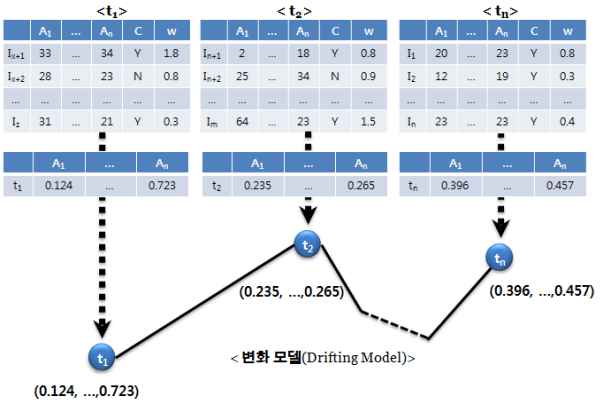


그림 1. 분할 영역 t에 따른 엔트로피 기반 공간 맵핑

그림 1과 같은 방법을 이용하여 전체 수집데이터를 분할하여 시간 t에 따른 엔트로피 테이블을 생성한다. 생성된 테이블은 n차원 공간에 사상하여 패턴 변화 모델로 나타낸다. 그림 2는 그림 1에서와 같은 방법으로 여러 가지 형태의 패턴 변화 모델들을 나타내고 있다. 이와 같이 공간상에 사상된 패턴 변화 모델은 그 흐름이 얼마나 유사한지 비교하여 유사 모델로 선정할 수 있다. 즉, 나와 비슷한 패턴 변화 모델이 있다면, 나의 모델에서 부족한 부분을 그 유사한 모델로부터 보충 할 수 있다. 그림 2에서 만약 User #3의 $t_4$ 이후를 예측한다고 가정할 때 , 가장 유사한 모델이 User #2 이므로, User #2의 $t_5$가 예측을 위해 사용 될 수 있다. 여기서 필요한 기술은 패턴 변화 모델간의 유사한 정도를 측정하는 방법이다. 이에 대해서는 다음 장에서 구체적으로 설명하겠다.

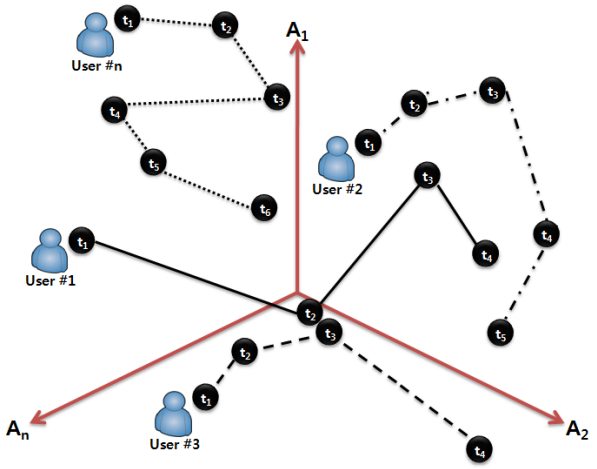


그림 2. 다양한 모델들의 n차원 공간 맵핑

### 3.2 모델 유사도

하나의 패턴 변화 모델은 공간상에 사상되었으며, 시간의 흐름에 따라 이동하는 하나의 경로(path)와 유사한 의미를 가지고 있다. 즉, 패턴 변화모델은 공간 지점의 위치와 그곳을 통과한 시각의 쌍의 집합으로 정의할 수 있다. 예를 들면 사용자가 $p_0$지점을 시간 0에 시작하여 $p_1$을 시간 30에 지났고 $p_2$를 40에, $p_3$을 65에 통과하여 $p_4$에 80에 도달했다면 모델 $P$는패턴 변화 모델은 수

$$P=\{(p_0,0),(p_1,30),(p_2,40),(p_3,65),(p_4,80)\}$$

으로 나타낼 수 있으며, 마찬가지로 임의의 다른 모델 $Q$는 아래와 같이 나타낼 수 있다.

$$Q=\{(q_0,0),(q_1,20),(q_2,50),(q_3,70),(q_4,75),(q_5,90)\}$$

또한, $P$와 $Q$의 지점에 대한 위치를 각각 점으로 표시하여 그림 3과 같이 나타낼 수 있다.

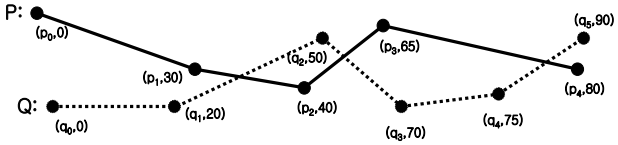


그림 3. 사용자 모델P와 Q의 공간 맵핑 사례

이때 시공간에서 발생한 두 패턴 변화 모델 $P$와 $Q$는 상호간에 얼마나 유사하다고 해야 할 것인가? 이 질문에 해답을 찾기 위해 다시 다음과 같은 세부적인 질문에 대답 할 수 있어야 할 것이다. “두 변화 모델이 얼마나 비슷한 시간을 가지고 이동하였는가?”, “시간에 따른 이동 변화가 얼마나 비슷한가?”, 그리고 “이동 변화의 유사정도와 함께 얼마나 방향이 일치 하는가?” 이 질문은 시간, 거리 그리고 방향 요소를 고려한다면 대답할 수 있을 것이다. 이를 위해 $P$의 임의의 지점과 $Q$의 임의의 지점간의 유사도를 정의하였다. 그 이유는 모델의 유사도를 두 모델이 얼마나 유사한 지점(노드, 모델)들이 포함되어 있는가를 이용하여 정의 할 수 있기 때문이다. 예를 들어 $P$의 한 지점 $(p_i, t_i)$와 $Q$의 한 지점 $(q_j, t_j)$의 유사도는 아래와 같이 정의할 수 있다.

$$nodeSim(p_i, q_j) = C_1^D \left[ \cos \frac{\theta}{2} \right]^{C_2} C_3^T$$

$C_1$, $C_2$, $C_3$ : 상수
$T$ : $|t_i - t_j|$
$D$ : 두 지점 $p_i$, $q_j$간의 거리
$\theta$ : $p_i$와 $q_j$의 사잇각

이를 바탕으로 모델 $P$와 $Q$의 유사도는 다음과 같이 정의하였다.

$$ModelSim(P,Q) = \max \sum_{i=0,j=0}^{m,n} nodeSim(p_i, q_j)$$

## 4. 실험 및 검증

패턴의 변화를 가지는 데이터에 기반을 둔 패턴 변화 모델을 생성하기 위해 인공 데이터를 생성하여 실험을 수행하였다. 데이터의 분포를 직관적으로 확인 할 수 있는 2차원 환경에서 인위적으로 데이터를 생성하고 패턴 변화 모델 간의 유사 정도를 확인하였다. 2개의 속성은 2차원 공간과 같은 의미를 가진다. x와 y 두 개 축을 가지는 환경에서 푸른색과 붉은색으로 구분하여 좌표를 지정하도록 하였다. 시간이 흐름에 따라 각 집단이 분리/이동을 진행하면서 네 모서리에 각각 분포하는 모습을 보인다. 이와 같은 방법으로 5가지의 패턴, 전체 10회의 변화를 가지는 데이터를 생성하였다.

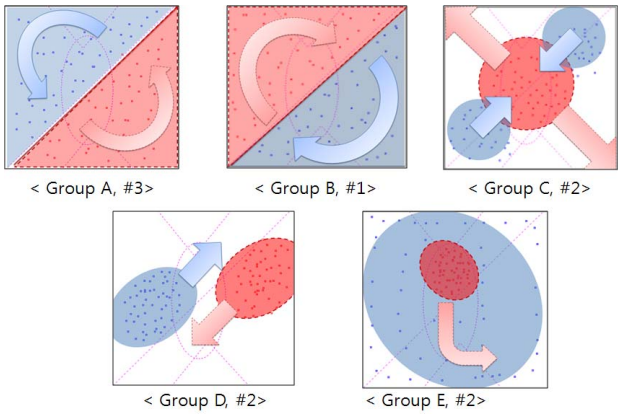


그림 4. 인공데이터 생성 사례

한 개의 데이터 집합은 10개의 변화 패턴을 가지도록 하였다. 그림 4에서 Group A는 중심의 사선을 기준으로 시계 반대방향으로 패턴이 이동하도록 3회 실시하였고, Group B는 중심의 사선을 중심으로 시계 방향으로 회전하는 집합을 1회 생성하였다. Group B를 생성한 이유는 Group A와 유사하지만 패턴의 변화가 상반된 모습을 보이도록 하여 얼마나 두 집합이 유사한지 비교하기 위함이다. 즉, Group A의 유사한 패턴 변화를 보이는 세 가지 집합과 Group B의 1가지 형태가 얼마나 연관성이 높은지를 비교하여 제안하는 방법이 적절한 결과를 보여주는지를 확인하였다. 이 외에 다양한 패턴을 모습을 비교하기 위해 Group C, D 그리고 E를 각각 2개씩 생성하여 유사정도를 비교하였다. 그림 5는 위의 10가지 다른 패턴 변화 모델 중에서 Group A-1, Group A-2 그리고 Group C-1의 변화 모습을 그래프로 표현한 것이다. 표 2는 앞서 제안한 방법을 이용하여 각 모델간의 유사도를 측정한 결과이다. 동일 그룹의 이동 패턴에 대해서는 0.5 이상의 유사도를 보이며, 그렇지 않은 경우 낮은 유사도를 보인다.

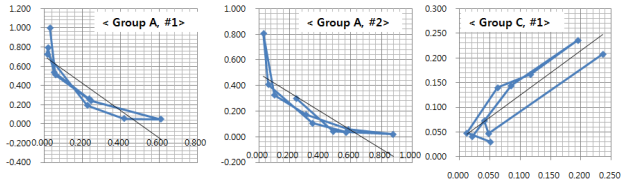


그림 5. Group A-1, Group A-2 그리고 Group C-1의 패턴 변화 모델의 변화 그래프

표 2. 패턴 변화 모델간 유사도 결과

| | | A | | | B | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 1 | 2 | 1 | 2 | 1 | 2 |
| A | 1 | 1.00 | 0.85 | 0.74 | 0.34 | 0.28 | 0.27 | 0.24 | 0.26 | 0.15 | 0.17 |
| | 2 | | 1.00 | 0.76 | 0.37 | 0.21 | 0.26 | 0.21 | 0.23 | 0.19 | 0.17 |
| | 3 | | | 1.00 | 0.35 | 0.16 | 0.18 | 0.21 | 0.25 | 0.19 | 0.20 |
| B | 1 | | | | 1.00 | 0.24 | 0.15 | 0.31 | 0.34 | 0.11 | 0.13 |
| C | 1 | | | | | 1.00 | 0.56 | 0.25 | 0.26 | 0.24 | 0.22 |
| | 2 | | | | | | 1.00 | 0.24 | 0.26 | 0.16 | 0.19 |
| D | 1 | | | | | | | 1.00 | 0.81 | 0.25 | 0.31 |
| | 2 | | | | | | | | 1.00 | 0.28 | 0.21 |
| E | 1 | | | | | | | | | 1.00 | 0.89 |
| | 2 | | | | | | | | | | 1.00 |

## 5. 결 론

본 논문은 패턴의 변화를 가지는 데이터의 분석을 위한 방법을 제안 하였다. 패턴 변화 데이터의 특성을 이용하여 패턴 변화 모델을 생성하고, 이 정보를 이용하여 다른 유사한 모델을 찾는데 사용하였다. 패턴의 변화를 고려한 패턴 변화 모델의 생성과 유사성 비교를 위한 실험에서는 2차원 공간에서 가상으로 데이터를 생성하여 실험하였다. 2차원 공간의 실험에서 유사한 패턴을 가지는 경우에 높은 모델 유사도 결과를 보여주었다.

## 참 고 문 헌

[1] C. C. Aggarwal, “Data Streams Models and Algorithms”, Springer US, pp. 1-7, 2007.
[2] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, “Models and issues in data stream systems”, ACM SIGMOD-SIGACT-SIGART Symposium on principles of database systems, 2002.
[3] A. Jain, “Statistical Mining in Data Streams”, Ph.D. Dissertation, University of California, Santa Barbara, 2006.
[4] 김진화, 민진영, “연속발생 데이터를 위한 실시간 데이터 마이닝 기법”, 한국경영과학회지, 제29권, 제4호, pp. 41-60, 2004.
[5] L. Golab, and M. T. Ozsu, “Issues in Data Stream Management”, SIGMOD Record, Vol. 32, No. 2, 2003.
[6] P. Domingos, and G. Hulten. “Mining high-speed data streams”, In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.