한국빅데이터학회지
제5권 제1호, 2020, pp. 147-154 https://doi.org/10.36498/kbigdt.2020.5.1.147

빅데이터 분석을 위한 파티션 기반 시각화 알고리즘

Partition-based Big Data Analysis and Visualization Algorithm

홍준기[1†]

배재대학교 컴퓨터공학과[1]

요 약

오늘날 빅데이터로부터 유의미한 결과를 도출하는 연구가 활발히 진행되고 있다. 본 논문에선 빅데이터의 데이터의 영역들을 파티션(partition)으로 설정하고 각 파티션들의 대표 값을 계산하여 변수들 사이의 상관관계를 분석 할 수 있는 파티션 기반 빅데이터 분석 알고리즘을 제안한다. 본 논문에선 파티션의 크기조절이 가능한 파티션 기반 빅데이터 분석 알고리즘의 파티션 크기 변화에 따른 시각화 결과를 비교분석하였다. 제안한 파티션 기반 빅데이터 분석 알고리즘을 검증하기 위해 의류 회사 'A'의 빅데이터를 분석하여 온도와 판매 가격 변화에 따른 상품의 판매량 변화를 분석하고 시각화하여 유의미한 결과를 얻을 수 있었다.

■ 중심어 : 빅데이터, 분석, 시각화, 파티션

Abstract

Today, research is actively being conducted to derive meaningful results from big data. In this paper, we propose a partition-based big data analysis algorithm that can analyze the correlation between variables by setting the data areas of big data as partitions and calculating the representative values of each partition. In this paper, the analyzed visualization results are compared according to the partition size of a proposed partition-based big data analysis (PBDA) algorithm that can control the size of the partition. In order to verify the proposed PBDA algorithm, the big data of 'A' is analyzed, and meaningful results are obtained through the analysis of changes in sales volume of products according to changes in temperature and sales price.

■ Keyword : Big Data, Analytics, Visualization, Partition

2020년 08월 05일 접수; 2020년 08월 20일 수정본 접수; 2020년 08월 25일 게재 확정.
† 교신저자 (jkhong@pcu.ac.kr)

Ⅰ. 서 론

오늘날 스마트 기기를 통한 물품 구매가 일상화됨에 따라 많은 온라인 쇼핑몰 기업들은 엄청난 양의 상품 판매 관련 빅데이터를 수집한다. 이처럼 수집된 빅데이터를 정제하고 분석하여 의미 있는 정보를 얻는 것은 기업의 상품판매 전략 수립과 효율적인 재고 관리를 위해 매우 중요하다. 따라서 최근 빅데이터 분석을 위한 많은 연구가 진행되었으며, 빅데이터 분석 기술 중 하나는 모든 수집된 빅데이터를 분석하는 것이 아닌 빅데이터의 일부분, 혹은 분할 계산하여 효율적으로 분석하는 것이다. 빅데이터를 분할하여 분석할 때 특정 구간, 혹은 범위를 의미하는 파티션을 사용하여 빅데이터를 분석하면 모든 데이터를 계산하지 않고 데이터의 통계적 특성을 추정 할 수 있는 장점이 있다. 이처럼 최근 전체 빅데이터를 분석하지 않고 랜덤 샘플 파티션을 사용하여 빅데이터를 분석하는 기술이 제안되었다 [1-2].

또한 최근 의류 상품과 관련된 많은 빅데이터 연구가 진행되고 있다 [3-5]. 소셜미디어의 빅데이터와 패션 상품의 시장 변화 관계를 분석한 연구가 진행되었으며 [6], 의류 관련 빅데이터를 수집하고 분석하여 판매량을 예측하는 연구도 활발히 진행되고 있다 [7-12]. 하지만 기존 파티션을 이용한 빅데이터 연구는 기업에서 수집된 빅데이터의 변수들 사이의 상관관계는 분석하지 않았으며 알고리즘의 성능만을 분석한 한계를 갖고 있다.

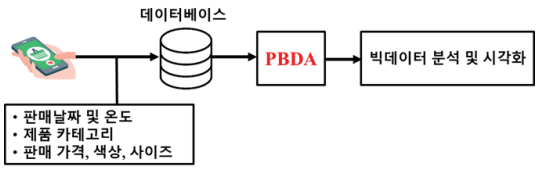
따라서 본 논문에선 빅데이터의 변수들 사이에서 유의미한 정보를 얻기 위해 파티션 기반 빅데이터 분석(Partition-based Big Data Analytics, PBDA) 알고리즘을 제안한다. 제안한 PBDA 알고리즘은 수집된 빅데이터를 파티션으로 분할하고 각 파티션의 대표 값을 계산하여 다른 변수들 사이의 상관관계를 효율적으로 분석할 수 있는 알고리즘이다. 또한 제안한 PBDA 알고리즘의 성능을 검증하기 위해 의류 회사 ‘A’에서 수집한 시계열 빅데이터를 분석하고 시각화하여 유의미한 결과를 도출 할 수 있었다.

Ⅱ. 제안한 빅데이터 분석 방법론

2.1 빅데이터 수집

본 연구에선 150만 이상 가입자를 보유하고 있는 국내 온라인 쇼핑몰 ‘A’에서 2014년 1월 1일부터 2018년 12월 31일까지 수집된 데이터를 사용하여 제안한 PBDA 알고리즘을 통해 빅데이터를 분석하였다.

온라인 쇼핑몰 ‘A’은 상품을 구매한 구매 날짜, 제품 카테고리, 판매 가격, 색상, 사이즈, 온도 정보를 실시간으로 데이터베이스(database, DB)에 저장한다. 온도 정보는 기상청의 국기기상종합정보 시스템인 ‘날씨누리’의 평균 온도를 수집하여 저장하며 그림 1은 제안한 PBDA 알고리즘을 포함한 전체 빅데이터 분석 순서도를 나타낸다.



〈그림 1〉 제안한 빅데이터 분석 순서도

제안한 PBDA 알고리즘을 사용하여 빅데이터를 분석하기 위해 DB에 저장된 상품 카테고리별 판매 날짜, 평균 온도, 강수량, 가격, 판매량 정보는 표 1과 같이 정제되어 CSV (comma-separated values) 형태로 저장된다.

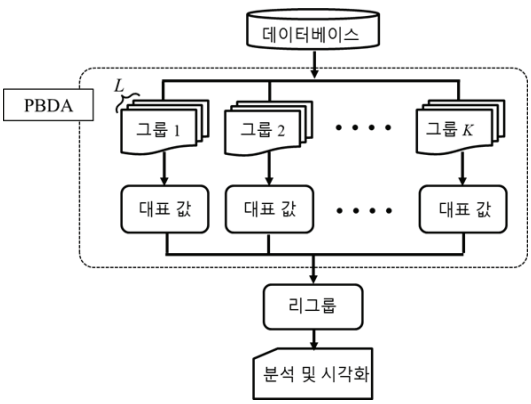
표 1에서 Date, avgTemp, rainFall, avgPrice, Sales는 각각 날짜, 평균 온도, 강수량, 판매 가격, 판매량을 나타낸다. 또한 각 상품의 데이터를 수집하고 상품별 판매량 특성을 분석하였다.

〈표 1〉 데이터 베이스 수집 예시

Date	avgTemp	rainFall	avgPrice	Sales
20140101	4	0	14,800	177
20140102	1.5	0	14,500	207
20140103	2.3	0	15,300	155
20140104	-0.1	0	15,200	122

2.2 PBDA

본 절에서는 제안한 PBDA 알고리즘을 설명한다. 제안한 PBDA 알고리즘의 순서도는 그림 2와 같이 나타낼 수 있다.



〈그림 2〉 빅데이터 분석 및 시각화 순서도

그림 2는 제안한 PBDA 알고리즘을 포함한 빅데이터 분석 순서도를 나타낸다. 데이터베이스에서 수집된 데이터는 L개의 원소로 이루어진 데이터 그룹들로 이루어지고 각 그룹들의 평균값은 각 그룹의 대표 값을 나타낸다. 이후 각 대표 값들만으로 빅데이터를 분석하고 시각화한다.

2014년 1월 1일부터 수집된 빅데이터는 순차적으로 저장되었다. 따라서 순차적으로 저장된 데이터는 L개의 원소로 이루어진 그룹들로 구분하고 대표 값을 계산하여 빅데이터를 효율적으로 분석하고 시각화한다. 2014년 1월 1일부터 수집된 날짜의 평균 온도 (average temperature, *AT*)은 다음과 같이 나타낼 수 있다.

$$\{AT_z\}_{z=1,\ldots,Z} \tag{1}$$

여기서 z는 2014년 1월 1일부터 2018년 12월 31일까지의 날짜 색인을 나타낸다. 예를 들어, 색인 z의 값 1, 2, 1825는 각각 2014년 1월 1일, 2014년 1월 21, 2018년 12월 31일을 의미한다.

낮은 온도에서 높은 온도로 변화할 때의 판매량 변화를 분석하기 위해 평균 온도 식 (1)은 식 (2)와 같이 오름차순으로 정렬한다.

$$\{AT_s\}_{s=1,\ldots,Z} \leftarrow sort\left(\{AT_z\}_{z=1,\ldots,Z}\right) \tag{2}$$

이후 파티션을 활용한 빅데이터의 효율적인 분석을 위해 오름차순으로 정렬된 배열을 L개의 원소로 이루어진 K개의 그룹으로 재배열한다.

다음은 정렬된 배열을 L개의 원소로 이루어진 K번째 그룹으로 재구성한 배열의 예시이다.

$$\{AT^{(k)}\} = \underbrace{[AT_{22}, AT_{436}, \ldots, AT_{1202}]}_{L} \tag{3}$$

또한 K번째 그룹의 동일한 날짜 색인의 판매량 (*SV*)은 다음 배열과 같이 나타낼 수 있다.

$$\{SV_{AT}^{(k)}\} = \underbrace{[SV_{22}, SV_{436}, \ldots, SV_{1202}]}_{L} \tag{4}$$

따라서 k 번째 배열의 평균 온도(*AT*)와 해당 판매량(*SV*)의 평균값은 다음과 같이 나타낼 수 있다.

$$\overline{AT^{(k)}} = \frac{1}{L}\sum_{s=1}^{L} AT_s^{(k)} \tag{5}$$

그러므로 *k*번째 그룹의 평균 온도 변화에 따른 판매량은 다음과 같이 나타 낼 수 있다.

$$\overline{SV_{AT}^{(k)}} = \frac{1}{L}\sum_{s=1}^{L} SV_s^{(k)} \quad (6)$$

또한 오름차순으로 정렬된 판매 가격의 *k*번째 그룹의 평균 판매량은 다음과 같이 나타 낼 수 있다.

$$\overline{P^{(k)}} = \frac{1}{L}\sum_{s=1}^{L} P_s^{(k)} \quad (7)$$

그러므로 *k*번째 그룹의 판매 가격 변화에 따른 판매량은 다음과 같이 나타 낼 수 있다.

$$\overline{SV_P^{(k)}} = \frac{1}{L}\sum_{s=1}^{L} SV_s^{(k)} \quad (8)$$

따라서, 그래프에 나타낼 총 *K*개의 평균 온도 배열(AT_f)과 판매량(SV_f^{AT}) 배열의 평균값 배열은 다음과 같이 나타낼 수 있다.

$$AT_f = \left[\overline{AT^{(1)}}, \overline{AT^{(2)}}, ..., \overline{AT^{(K)}}\right] \quad (9)$$

$$SV_{AT_f} = \left[\overline{SV_{AT}^{(1)}}, \overline{SV_{AT}^{(2)}}, ..., \overline{SV_{AT}^{(K)}}\right] \quad (10)$$

또한 총 *K*개의 판매 가격 배열(P_f)과 판매량(SV_f) 배열의 평균값 배열은 다음과 같이 표현 할 수 있다.

$$P_f = \left[\overline{P^{(1)}}, \overline{P^{(2)}}, ..., \overline{P^{(K)}}\right] \quad (11)$$

$$SV_{P_f} = \left[\overline{SV_P^{(1)}}, \overline{SV_P^{(2)}}, ..., \overline{SV_P^{(K)}}\right] \quad (12)$$

다음 절에서는 파티션의 원소개수 L 변화에 따른 평균 온도와 판매 가격 변화에 따른 상품의 판매량 변화를 분석하고 시각화 한다.

Ⅲ. PBDA를 통한 상품의 판매량 분석

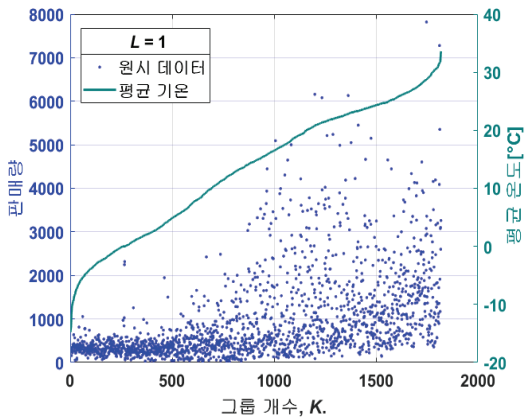
본 장에서는 2장에서 제안한 PBDA 알고리즘을 통해 *L*값 변화와 평균온도, 판매 가격 변화에 따른 반팔 티셔츠, 반바지, 신발의 판매량 변화를 시각화하고 분석한다.

3.1 반팔 티셔츠 판매량 분석결과

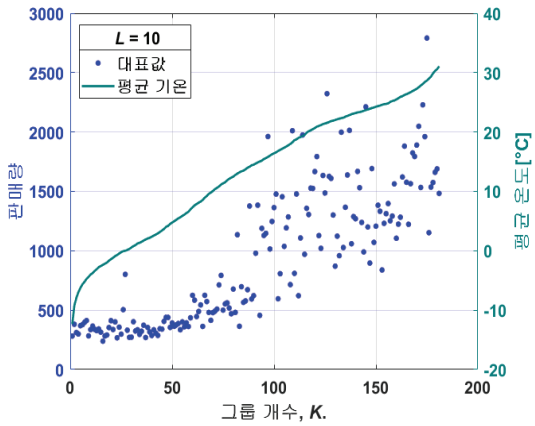
이번 절에서는 2장에서 제안한 PBDA 알고리즘을 통해 분석한 빅데이터의 *L*값 변화에 따른 반팔 티셔츠와 겨울 자캣의 판매량 변화를 분석하고 시각화한다.

그림 3의 x축, 왼쪽 y 축, 오른쪽 y 축은 각각 그룹의 개수, 판매량, 평균 온도를 나타낸다. 그림 3 (a)는 오름차순으로 정렬된 평균온도 변화에 따른 반팔 티셔츠의 판매량을 나타낸 그래프이다. *L*=1 일 때, 즉 파티션 분석기법을 사용하지 않고 오름차순으로 정렬된 원시 데이터의 평균온도 변화에 따른 반팔 티셔츠의 판매량 변화를 나타낸 결과이다. 그림 3(a)에 따르면 파티션 기법을 사용하지 않았을 때 원시데이터로부터 유의미한 분석 결과를 얻기는 매우 어려운 것을 확인 할 수 있다.

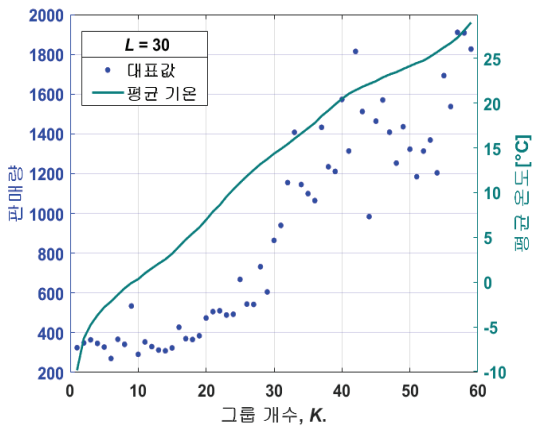
그림 3(b)는 파티션의 크기, 즉 *L*의 크기를 10으로 설정하고 각 그룹의 대표값을 사용하여 평균 온도 변화에 따른 반팔 티셔츠의 판매량 변화를 나타낸 그림이다. 그림 3(b)에서 확인 할 수 있듯이, 평균 온도가 상승함에 따라 반팔 티셔츠의 판매량도 증가하는 것을 확인할 수 있다. 또한 그림 3(c)는 파티션의 크기를 *L*=30으로 설정했을 때의 결과 그래프이며 파티션의 크기가 *L*=10일 때보다 더욱 명확히 판매 가격이 상승함에 따라 반팔 티셔츠의 판매량이 증가하는 것을



(a)

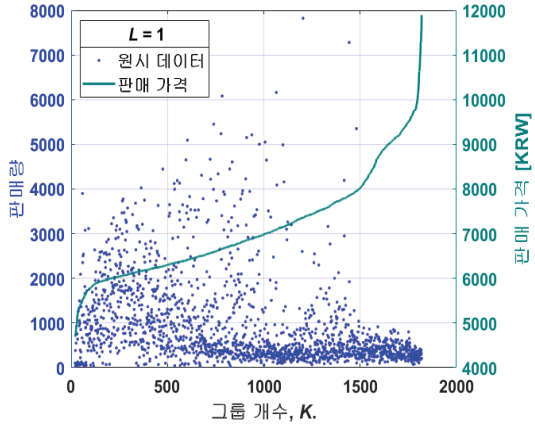


(b)

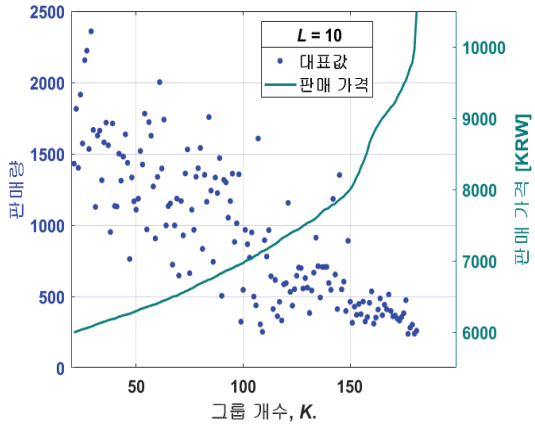


(c)

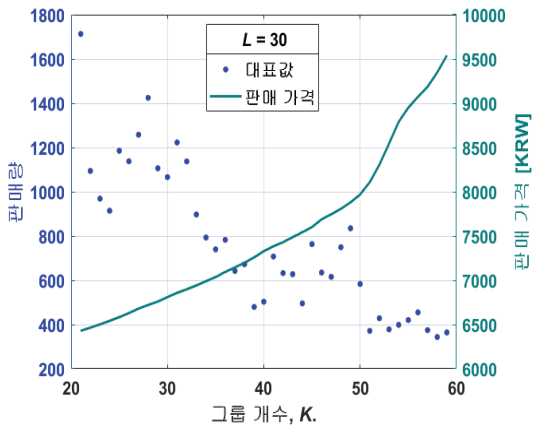
〈그림 3〉 평균 온도 변화에 따른 반팔 티셔츠의 판매량 분석 결과 (a) L=1, (b) L=10, (c) L=30.



(a)

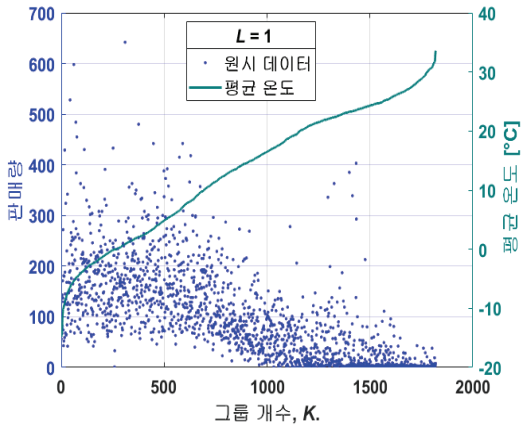


(b)

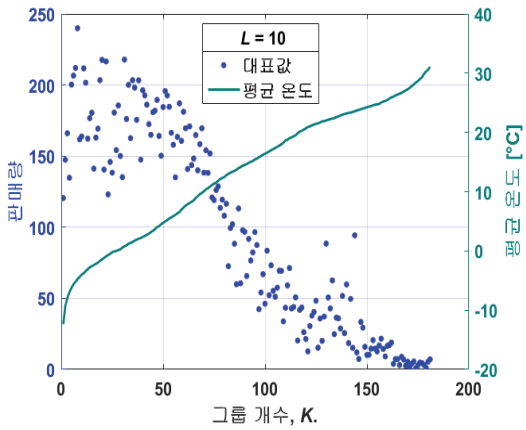


(c)

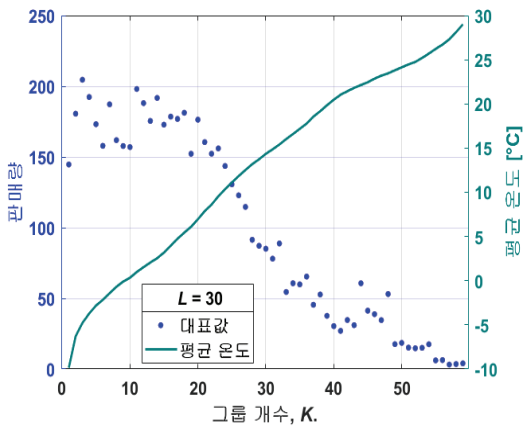
〈그림 4〉 판매 가격 변화에 따른 반팔 티셔츠의 판매량 분석 결과 (a) L=1, (b) L=10, (c) L=30.



(a)

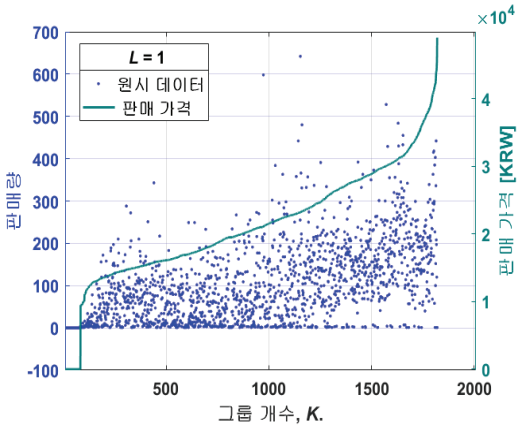


(b)

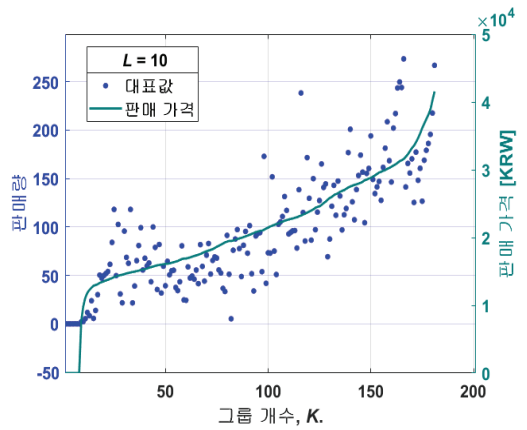


(c)

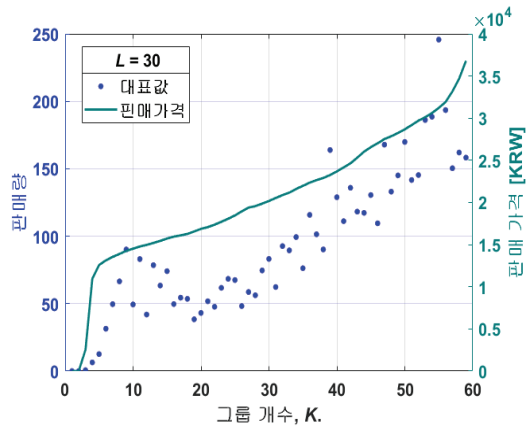
〈그림 5〉 평균 온도 변화에 따른 재킷의 판매량 분석 결과 (a) L=1, (b) L=10, (c) L=30.



(a)



(b)



(c)

〈그림 6〉 판매 가격 변화에 따른 재킷의 판매량 분석 결과 (a) L=1, (b) L=10, (c) L=30.

확인 할 수 있다. 또한, 그림 4는 가격 변화에 따른 반팔티셔츠 판매의 판매량을 나타낸 그래프이다. 그림 4(a)에 따르면 그림 3(a) 결과와 마찬가지로 제안한 PBDA 알고리즘을 사용하지 않고 오름차순으로 정렬된 원시 데이터로는 유의미한 결과를 얻기 어려운 것을 확인 할 수 있다.

하지만 제안한 PBDA를 사용하여 파티션의 크기를 L=10, L=30으로 증가시킬수록 판매 가격이 증가할수록 반팔 티셔츠의 판매량이 감소하는 유의미한 결과를 얻을 수 있는 것을 확인 할 수 있다.

3.2 재킷의 판매량 분석결과

본 절에선 높은 온도에서 많이 판매되는 반팔 티셔츠와는 반대의 성격을 갖는 재킷의 판매량을 분석하여 제안한 PBDA 알고리즘의 성능을 평가하였다. 다음 그림 5는 제안한 PBDA 알고리즘을 사용한 평균 온도 변화에 따른 재킷의 판매량 변화를 분석한 그래프이다.

그림 5(a)에서 확인 할 수 있듯, 파티션의 크기가 1인 경우 수집된 빅데이터로부터 유의미한 결과를 얻을 수 없는 것을 확인하였다. 하지만 제안한 PBDA 알고리즘을 사용하여 파티션의 크기를 L=10, L=30으로 증가시키고 각 그룹의 대표값으로 분석하였을 때 평균 온도가 상승할수록 재킷의 판매량이 감소하는 분석 결과를 더욱 명확하게 확인 할 수 있었다.

그림 6은 판매가격 변화에 따른 재킷의 판매량을 나타낸 그래프이다. 그림 6에 따르면 파티션의 크기가 상승 할수록 재킷의 판매량이 상승하는 것을 확인 할 수 있다. 이와 같은 재킷의 판매량 결과는 여름에 판매되는 반팔과는 다르게 가격이 상승하더라도 보온성과 품질이 뛰어난 제품을 구매하기 때문이다.

Ⅳ. 결 론

본 연구에서는 제안한 빅데이터를 효율적으로 분석하기 위해 파티션을 사용한 PBDA 알고리즘을 통해 파티션 크기 변화에 따른 반팔과 재킷의 온도와 판매가격 변화에 따른 판매량 변화를 분석하였다. 2014년 1월 1일부터 온라인 쇼핑몰 'A'에서 수집된 빅데이터를 사용하여 파티션 크기 변화에 따른 상품의 판매량을 시각화하고 분석하였다. 빅데이터 분석 결과에 따르면 제안한 PBDA 알고리즘의 파티션 기법을 사용하지 않고 빅데이터를 시각화하였을 때 유의미한 결과를 확인하기 어려웠지만 파티션의 크기를 확대하여 빅데이터로부터 유의미한 판매량 분석 결과를 얻을 수 있었다.

분석 결과에 따르면, 반팔 티셔츠의 판매량은 온도가 상승하고 가격이 하락함에 따라 증가하는 것을 확인하였으며 재킷의 판매량은 온도가 하락하고 가격이 상승함에 따라 증가하는 것을 확인 할 수 있었다. 따라서 제안한 PBDA 알고리즘을 통해 모든 빅데이터를 사용하여 분석하는 것보다 훨씬 효율적으로 시각화 할 수 있는 것을 확인 할 수 있었다.

따라서 제안한 PBDA 알고리즘을 통해 시계열 빅데이터의 변수의 상관관계를 2D 그래프를 통해 효율적으로 분석할 수 있었다. 또한 추후 더욱 정교한 대표 값을 계산하는 알고리즘을 제안한 PBDA 알고리즘에 적용하여 더욱 정확한 변수간의 상관관계를 분석할 수 있는 연구를 진행할 예정이다.

참 고 문 헌

[1] M, Vojnovic, F. Xu, J. Zhou, "Sampling based range partition methods for big data analytics." Microsoft Res., Microsoft Corp., Redmond, WA,

USA, Tech. Rep. MSR-TR-2012-18

[2] L. Wu, R. J. Barker, M. A. Kim, A. K. A. Ross, “Navigating big data with high-throughput, energy-efficient data partitioning.” ACM SIGARCH Comput. Arch. News, Vol. 41, No. 3, 249-260, 2013.

[3] S. Jain, J. Bruniaux, X. Zeng, and P. Bruniaux, “Big Data in Fashion Industry,” IOP Conference Series: Materials Science and Engineering, Vol. 254, No. 15, 2017.

[4] D. Øivind and T. Stenheim, “Big Data Viewed Through the Lens of Management Fashion Theory,” Cogent Business & Management, Vol l, No. 3, 2016.

[5] New York Times, The Age of Big Data(2012). https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[6] T. Choi, “Incorporating Social Media Observations and Bounded Rationality into Fashion Quick Response Supply Chains in the Big Data Era,” Transportation Research Part E: Logistics and Transportation Review, Vol. 114, pp. 386-397, 2018.

[7] S. Thomassey, “Sales Forecasts in Clothing Industry: The Key Success Factor of the Supply Chain Management,” International Journal of Production Economics, Vol. 128, No. 2, pp. 470-483, 2010.

[8] W. K. Wong and Z. X. Guo, “A Hybrid Intelligent Model for Medium-term Sales Forecasting in Fashion Retail Supply Chains using Extreme Learning Machine and Harmony Search Algorithm,” International Journal of Production Economics, Vol. 128, No. 2, pp. 614-624, 2010.

[9] S. Ren, T. Choi, and N. Liu, “Fashion Sales Forecasting with a Panel Data-based Particle-filter Model,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 45, No. 3, pp. 411-421, 2015.

[10] K. Au, T. Choi, and Y. Yu, “Fashion Retail Forecasting by Evolutionary Neural Networks,” International Journal of Production Economics, Vol. 114, No. 2, pp. 615-630, 2018.

[11] Y. Ni and F. Fan, “A Two-stage Dynamic Sales Forecasting Model for the Fashion Retail,” Expert Systems with Applications, Vol. 38, No. 3, pp. 1529-1536, 2011.

[12] N. Liu, S. Ren, T.-M. Choi, C.-L. Hui, and S.-F. Ng, “Sales Forecasting for Fashion Retailing Service Industry: A Review,” Mathematical Problems in Engineering, Vol. 2013, pp. 1-9, 2013.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1G1A1100225).

저 자 소 개

홍 준 기(Jun-Ki Hong)

·2010년 11월: Carleton University 컴퓨터 시스템 공학과 (학사)
·2010년 9월~2017년 2월: 연세대학교 전기전자공학과 (박사)
·2016년 8월~2017년 7월: 한국정보통신기술협회(TTA) 선임연구원
·2017년 8월~2020년 2월: 영산대학교 전기전자공학과 조교수
·2020년 3월~현재: 배재대학교 컴퓨터공학과 조교수
·관심분야: IoT, 빅데이터, 인공지능, 항공체, 차세대 통신 등