

빅데이터를 기반으로 한 20대의 패션 트렌드 분석

양윤정 · 엄병용 · 홍성엽 · 유동희

부산가톨릭대학교

An Analysis of Twenties Fashion Trend based on Big data

Yoon-jung Yang · Byung-yong Um · Sung-yub Hong · Donghui Yu

Catholic University of Pusan

E-mail : dbswjd4980@naver.com

요 약

최근 빅데이터의 등장으로 그에 따른 활용이 굉장히 광범위 해지고 있다. 빅데이터를 기반으로 한 행정자치 및 교통통제 서비스가 사용되고 있으며 앞으로도 빅데이터를 활용한 많은 서비스들을 사용할 수 있을 것이다. 이에 본 논문에서는 20대들이 자주 찾는 매장의 판매 데이터나 온라인 쇼핑몰의 검색 및 조회 순 등의 빅데이터를 정의하고 이를 활용한 20대 패션 트렌드 분석 방법을 제안하고 최적의 상품 진열 방법 등을 제시하여 판매율을 제고시키고자 한다.

키워드

빅데이터, 매장관리, 매출, 유행, 분석

1. 서 론

인터넷과 모바일기기의 이용이 생활화되면서 다양한 종류의 데이터들이 기하급수적으로 증가하고 있다. 기존의 데이터베이스로는 이를 처리하는데 한계가 있어 빅데이터가 등장하게 되었다.

빅데이터란 기존의 데이터베이스가 저장, 관리, 분석할 수 있는 범위를 초과한 규모의 데이터를 말하며 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형 데이터를 의미한다[1].

본 논문에서는 의류관련 데이터를 수집하여 빅데이터 기반으로 우리나라 20대들의 패션트렌드를 분석하는 방법을 제시한다.

1) 데이터 수집

데이터 수집은 크게 오프라인 매장데이터와 온라인 쇼핑몰데이터, 패션블로그/사이트 데이터로 나누어 수집한다.

오프라인 매장 데이터에서는 해당 브랜드로부터 전체 의류 판매데이터와 회원별 판매데이터를 수집한다. 전체 판매데이터를 통해서 한 달 평균 구입자 수를 계산하여 매장의 순위를 측정하는데 사용한다. 또한 회원별 판매데이터는 회원 중 20대들이 구매하는 의류가 무엇인지 파악하는데 사용한다.

그림 1과 그림 2는 오프라인 매장에서부터 수집한 자료의 예시를 나타낸 것이다. 온라인 쇼핑몰 데이터에서는 쇼핑몰 접속자 데이터와 전체 판매데이터, 회원데이터를 수집한다.

II. 트렌드 분석절차

본 논문에서 제안하는 트렌드 분석절차는 데이터 수집, 데이터 전처리 및 저장, 군집분석, 연관성분석으로 총 4단계를 걸쳐 분석결과를 도출하도록 구성되어 있다. 이러한 단계를 거쳐 20대들에게 인기있는 의류와 쇼핑몰, 브랜드 순위, 국내외 스트리트 패션, 드라마 속 패션에 대한 정보를 제공하고, 각 브랜드와 쇼핑몰에는 인기상품 정보를 이용하여 효과적인 진열배치를 위한 정보를 제공할 것이다.

품목	색상	월일	거래처	입고				출고				재고							
				55	66	77	88	계	55	66	77	88	계	55	66	77	88	계	
G08-J-001 BK			직판재고	320		230			550					320		230		550	
G08-J-001 BK		03월 29일	강남매장						2	3			5	338	27	260		625	
G08-J-001 BK		03월 30일	인천매장	1				1						338	25	260		623	
G08-J-001 GR			직판재고	600				600		4			4	600				600	
G08-J-001 GR		03월 29일	강남매장											610	16	30		656	
G08-J-001 GR		03월 30일	인천매장		21			21						610	38	30		677	
품목별합계				61	71	60		192	3	9			12						
G08-J-006 BK			직판재고	90		120		150	360					90		120		360	
G08-J-006 BK		03월 29일	명동매장						2	3	4	5	14	198	217	446	185	1046	
G08-J-006 BK		03월 30일	인천매장						2					198	210	446	185	1041	
품목별합계				111	220	332		40	703	3	10	4	5	22					
합계				172	291	392		40	895	6	19	4	5	34	176	282	738	185	2371

그림 1. 오프라인 전체 의류 판매데이터

회원번호	성별	생년월일	구입월일	품목	색상	구입매장
534A897	여자	930515	03월 29일	G08-J-001	BK	강남매장
795G369	남자	871005	03월 29일	G08-H-001	WH	인천매장
153G657	남자	910628	03월 29일	D10-J-215	NB	강남매장
205E315	여자	850305	03월 30일	G08-J-001	BK	명동매장
340C884	여자	901209	03월 30일	G21-J-122	PN	인천매장
591B507	여자	960407	03월 30일	R15-F-853	PN	인천매장

그림 2. 오프라인 회원별 판매데이터

그림 3과 같이 쇼핑물 접속자 데이터에서는 한 달 평균 접속자를 파악하여 쇼핑물 순위 측정에 사용한다. 전체 판매데이터와 회원데이터를 이용하여 20대들이 주로 구입하는 의류를 파악한다.

#	IP Address	Time	Method/ URL/ Protocol	Status	Size	Referred
1	192.138.58.7	[29/Mar/2014:23:57:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-
2	185.168.58.8	[29/Mar/2014:23:57:46 -0500]	"GET A.html HTTP/1.0"	200	2050	A.html
3	192.168.58.7	[29/Mar/2014:23:57:50 -0500]	"GET A.html HTTP/1.0"	200	4130	-
4	192.165.58.4	[29/Mar/2014:23:58:01 -0500]	"GET A.html HTTP/1.0"	200	5096	B.html
5	192.118.58.7	[29/Mar/2014:23:58:07 -0500]	"GET A.html HTTP/1.0"	200	3290	-
6	192.168.58.8	[29/Mar/2014:23:59:07 -0500]	"GET A.html HTTP/1.0"	200	2050	A.html
7	192.168.32.9	[29/Mar/2014:23:59:09 -0500]	"GET A.html HTTP/1.0"	200	8140	L.html
8	192.168.58.1	[29/Mar/2014:03:09:13 -0500]	"GET A.html HTTP/1.0"	200	1820	A.html
9	135.168.58.3	[29/Mar/2014:03:09:58 -0500]	"GET A.html HTTP/1.0"	200	2270	F.html
10	192.168.58.5	[30/Mar/2014:00:01:02 -0500]	"GET A.html HTTP/1.0"	200	9436	C.html
11	192.157.58.2	[30/Mar/2014:00:01:09 -0500]	"GET A.html HTTP/1.0"	200	7220	B.html
12	192.168.36.7	[30/Mar/2014:00:02:55 -0500]	"GET A.html HTTP/1.0"	200	3290	-
13	192.168.23.6	[30/Mar/2014:00:07:09 -0500]	"GET A.html HTTP/1.0"	200	1680	A.html

그림 3. 온라인 쇼핑물 접속 기록

또한 회원데이터를 이용하여 회원들 중 20대의 비율을 계산하여 20대에게 인기 있는 쇼핑물의 순위를 측정한다.

오프라인 매장데이터와 달리 온라인 쇼핑물 데이터에서 회원별 판매데이터를 수집하지 않는 이유는 대부분의 온라인 쇼핑물은 회원가입을 전제로 하여 주문을 할 수 있기 때문에 판매데이터에 회원의 정보가 포함되어 있다.

패션블로그/사이트로부터는 사람들이 많이 찾아보는 의류를 파악하기 위해 사전에 패션블로그와 사이트를 지정해 놓는다. 그림 4와 그림 5는 국내외 스트리트 패션과 드라마 속 연예인의 패션에 관한 정보가 있는 사이트에 대한 사진이다.

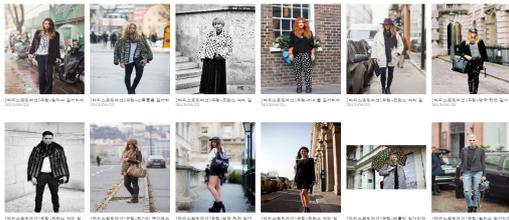


그림 4. 국내외 스트리트 패션사이트



그림 5. 드라마 속 연예인 패션사이트

본 논문에서는 Flume을 사용하여 패션블로그와 사이트의 데이터를 수집할 것이다. Flume이란 분산 환경에서 대량의 로그 데이터를 효과적으로 수집, 처리하는 신뢰할 수 있는 서비스이다. 또한, Flume은 장애에 쉽게 대처가 가능하며, 로그 유실에 대한 신뢰수준에 맞게 상황을 변경할 수 있을 뿐만 아니라 장애 발생시에 다양한 복구 메커니즘이 제공한다.[2]

Flume을 이용해 사람들이 패션블로그나 사이트의 게시물을 클릭하였을 때 발생하는 데이터를 수집하여 방문자들이 가장 많이 본 게시물의 사진이나 동영상상을 파악한다.

2) 데이터 저장 및 전처리

이렇게 수집한 데이터는 Hive를 이용하여 저장한다. Hive는 분산 스토리지에 저장되어진 대용량 데이터의 조회 및 관리를 용이하게 해 주는 툴이다. HiveQL은 SQL과 흡사한 언어를 사용하여 데이터를 조회하는 구조를 가지고 있다[3]. 데이터 입력 형식은 표 1과 같다.

입력한 데이터를 활용하여 Hive에서 필요한 내용을 질의 후 Hadoop을 통해 MapReduce환경에서 처리한다.

표 1. Hive입력 데이터 형식

Table	Column	Type	변수
매장 판매데이터	Item	FLOAT	품목
	Color	SMALLINT	색상
	Size	SMALLINT	사이즈
	Account	STRING	판매처
	Volume	INT	판매수량
회원별 매장 판매데이터	Member_N	FLOAT	회원번호
	Member_B	FLOAT	생년월일
	Gender	SMALLINT	성별
	Item	STRING	품목
	Color	SMALLINT	색상
쇼핑물 판매데이터	Volume	INT	구입수량
	Item	STRING	품목
	Color	SMALLINT	색상
	Size	SMALLINT	사이즈
	Member_N	FLOAT	회원번호
	Member_B	FLOAT	생년월일
쇼핑물 접속자 기록	Gender	SMALLINT	성별
	Volume	INT	구입수량
	Month	SMALLINT	월
	Avg_Access	INT	평균 접속자 수
	Member_N	FLOAT	회원번호
쇼핑물 회원 데이터	Member_B	FLOAT	생년월일
	Gender	SMALLINT	성별

3) 군집분석

군집분석이란, 데이터의 유사성을 이용하여 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 분석법을 말한다[4].

군집분석은 크게 계층적 군집분석과 비계층적 군집분석으로 나뉜다. 계층적 군집분석에는 병

합방법, 최단연결법, 최장연결법이 있고 비계층적 군집분석에는 K-평균 군집, 가우스 혼합모형, 자기조직화 지도 등이 있다.[4]

본 논문에서는 오프라인 회원별 판매 데이터와 온라인 판매데이터, 회원데이터를 바탕으로 20대들이 주로 구입하는 의류를 분석하기 위하여 군집분석을 사용할 것이다.

군집화를 하기위한 변수로는 인구통계적 변수를 사용한다. 인구통계적 변수에는 성별, 나이, 거주지, 직업, 소득, 등이 있다. 이러한 변수들 중에서 나이를 기준으로 군집화하여 20대들에게 인기 있는 의류를 분석한다.

4) 연관성분석

연관성분석이란, 데이터 안에 존재하는 항목간의 연관규칙(association rule)을 발견하는 과정을 말한다. 여기서 연관규칙은 상품을 구매하거나 서비스를 받는 등의 일련의 거래나 사건들의 연관성에 대한 규칙을 말한다[4].

연관성분석의 측도로는 지지도와 신뢰도가 있다. 예를 들어 설명하자면 두 품목 A 와 B의 지지도는 전체 거래품목 중 품목 A와 B가 동시에 포함하는 거래의 비율을 말하고, 신뢰도는 품목 A를 포함하는 거래 중 품목 A와 B를 동시에 포함하는 거래의 비율을 말한다.[4]

본 논문에서는 연관성분석을 이용하여 오프라인 매장과 온라인 쇼핑몰 의류 배치의 효율성을 향상시키는데 도움을 줄 것이다.

5) 예측 결과

본 논문에서 제시한 분석 방법에 따라 결과를 예측해보면 우선적으로 구입자에게 20대 인기 상품과 브랜드 순위, 매장순위, 쇼핑몰 순위를 제공할 수 있다.

그리고 현재 국내외로 유행하는 스트리트 패션이나 드라마 속 연예인의 패션에 관한 정보도 제공할 수 있다.

또한 브랜드와 각 매장에는 브랜드별 인기상품과 효과적인 진열배치를 위한 정보도 제공할 수 있다.

III. 결 론

현재까지는 빅데이터를 기반으로 하여 의류의 판매 효율성에 대해 연구한 논문은 없었다. 그런 점에서 본 논문은 최근 이슈화되고 있는 빅데이터를 기반으로 구입자에게 더 많은 서비스와 정보, 편리함을 제공한다는 점에서 의의를 가진다.

또한 이러한 분석과정을 통해 얻은 결과를 이용하여 구입자에게 제공함으로써 고객의 만족도를 향상시키고 기업의 이익과 이미지 쇄신에도 큰 영향을 미칠 것으로 예상된다.

추후 Hadoop과 MapReduce환경에서 실제로 데이터를 수집하고 데이터의 저장 및 전처리 후

비계층 군집분석 중 K-평균 군집분석과 지지도를 측도로 한 연관성분석을 적용하여 결과를 도출해 보는 것도 시도하여 볼 것이다.

참고문헌

- [1] Kim BS, "Understanding Key Factors of Users' Intentions to Repurchase and Recommend Digital Items in Social Virtual Worlds," *Cyberpsychology, Behavior, and Social Networking* 2012, Vol. 15, No. 10, pp. 543-550. doi: 10.1089/cyber.2012.0128.
- [2] 한수현, 빅데이터 활용을 위한 데이터 수집 도구 평가에 관한 연구, pp. 15, 2013.12
- [3] Apache TM, "Apache TM Hive", <http://hive.apache.org>
- [4] 박창이, 김진석, R을 이용한 데이터마이닝, pp. 79, 66, 81-89, 68-69 2008.12