JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 8, pp. 65-72, August 2020
https://doi.org/10.9708/jksci.2020.25.08.065

Developing a Big Data Analysis Platform for Small and Medium-Sized Enterprises

Hyeon Gyu Kim*

*Associate Professor, Div. of Computer Science and Engineering, Sahmyook University, Seoul, Korea

[Abstract]

Big data analysis is widely used in applications such as finance and communication, whose market size is growing rapidly every year. Nevertheless, it is rarely used by SMEs (small and medium-sized enterprises) since the existing services are not fully customized for them while being offered at high price. To resolve this, we develop and propose a new platform to provide big data analysis services specialized for SMEs in this paper. First, we compare existing work discussing social big data analysis, and extract service features necessary to help their marketing effectively. Then, we present a prototype system implementing the extracted features, and discuss technical issues needed to develop a complete system which are obtained from the prototype implementation.

▸ **Key words**: Big data, Social reviews, Morphological analysis, Noise review filtering, SMEs (small and medium-sized enterprises)

[요 약]

금융, 통신 등의 응용 분야에서 빅데이터는 광범위하게 활용되고 있으며, 빅데이터 분석 시장은 해마다 크게 성장하고 있다. 이에 반해 소상공인들의 빅데이터 활용 실적은 저조하며, 이는 기존 시스템이 소상공인들의 여건을 충분히 반영하지 못하는 동시에 서비스 이용 가격 역시 높다는 점에 기인한다. 이를 해결하기 위한 노력의 일환으로, 본 논문에서는 소상공인에 특화된 빅데이터 분석 서비스를 제공하는 새로운 플랫폼을 개발, 제안한다. 먼저 소셜 빅데이터 분석과 관련한 기존 연구들을 비교하고, 소상공인의 마케팅을 돕기 위해 필요한 서비스 지표들을 추출한다. 다음으로 도출된 지표들을 구현한 프로토타입 시스템을 소개하고, 구현을 통해 얻어진 시스템 완성에 필요한 기술적인 이슈들을 논의한다.

▸ **주제어**: 빅데이터, 소셜 리뷰, 형태소 분석, 노이즈 리뷰 필터링, 소상공인

• First Author: Hyeon Gyu Kim, Corresponding Author: Hyeon Gyu Kim
*Hyeon Gyu Kim (hgkim@syu.ac.kr), Div. of Computer Science and Engineering, Sahmyook University
• Received: 2020. 07. 03, Revised: 2020. 08. 05, Accepted: 2020. 08. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

빅데이터란 3V(Volume, Velocity, Variety) 속성을 지니는 데이터로 정의될 수 있으며, 대표적인 사례로 신용카드 트랜잭션, 휴대폰 전화 및 문자 내역, SNS 피드 및 블로그 리뷰 등을 포함한다[1, 2]. 국내에서는 금융, 보험, 부동산, 통신 등의 응용에서 고객의 요구를 추출하고 반영하기 위해 빅데이터 분석을 수행하는 사례가 증가하고 있으며, 이와 관련한 국내 시장 규모는 2019년 기준으로 1조 6744억원 정도로 추산되고 있다[3].

이에 반해 빅데이터 분석은 자본력 있는 대기업 및 공기업 등에 의해 주로 이용되고 있으며, 자금 여력이 부족한 소상공인들은 제대로 활용하지 못하고 있는 실정이다. 가장 큰 이유는 기존 빅데이터 분석 솔루션의 가격이 소상공인이 감당할 수 있는 수준을 넘어선다는 것이다. 예를 들어 소셜 빅데이터 분석 관련 국내 대표적인 선두주자 중 하나인 메조미디어 티버즈(TIBUZZ)[4]의 경우, 월별 사용료가 600만원(VAT 별도)이며 기본 계약 기간이 최소 3개월이므로, 소상공인이 해당 솔루션을 이용하기에는 현실적으로 어려운 실정이다.

비용 문제를 해결하기 위해 텍스톰(Textom)[5]과 같은 소셜 리뷰 수집과 키워드 추출 서비스를 무료로 제공하는 도구를 이용할 수 있으나, 이 경우 소상공인들이 노이즈 리뷰 필터링을 직접 수행해야 하는 번거로움이 있다. 노이즈 리뷰란 주어진 키워드와 연관성이 없는 소셜 리뷰를 의미하며, 해당 리뷰가 제대로 걸러지지 않을 경우 유의미한 분석 결과를 얻기 어려워진다. 따라서 전문성이 부족한 소상공인들이 노이즈 리뷰 필터링을 수행하기는 사실상 어려우며, 설령 수행한다고 하더라도 리뷰에서 추출된 키워드만을 볼 수 있어, 업체의 현황 분석 및 마케팅에 효과적으로 활용하기에는 한계가 있다.

그럼에도 불구하고 소상공인들에 대한 빅데이터 분석은 경쟁력 향상을 위해 필수라고 할 수 있다. 본 연구진이 ㈜온굿플레이스[6]의 의뢰를 통해 경상남도 통영시 및 거제시 상인들을 대상으로 조사를 수행한 결과, 상인들이 인지하고 있는 내용과 빅데이터 분석 내용에 다음과 같은 차이점이 존재하는 것을 확인하였다.

○ 인기메뉴: 상인들이 생각하는 대표 메뉴와 고객들이 선호하는 메뉴가 상이함
○ 경쟁상대: 상인들이 생각하는 경쟁 상대와 빅데이터 분석을 통해 정량화한 경쟁 상대가 상이함 (다수의 상인들이 자신보다 훨씬 인지도가 높은 상대를 경쟁 상대로 인지함)

본 논문에서는 소상공인들의 마케팅 경쟁력 향상을 지원하기 위한 빅데이터 분석 플랫폼의 개발 현황과 이슈에 대해 소개한다. 먼저 2장에서는 소상공인 대상 빅데이터 분석과 관련한 기존 연구들을 소개한다. 3장에서는 소상공인을 대상으로 한 빅데이터 분석이 기존의 분석 방법과 어떤 차이점이 있는지 설명한다. 4장에서는 해당 차이점을 기반으로 소상공인들의 마케팅에 도움이 될 수 있는 서비스 지표들을 추출한다. 5장에서는 추출된 지표를 구현한 프로토타입 시스템의 구조와 실행 예들을 소개하고, 구현을 통해 얻어진 기술적인 이슈들을 논의한다. 마지막으로 6장에서는 결론 및 추후 연구 방향 제시로 마무리한다.

II. Related Work

빅데이터 분석을 활용하여 소상공인의 마케팅을 지원하고자 시도한 접근 방법은 최근 들어 논의가 진행되고 있으며, 어떤 데이터를 활용하느냐와 관련하여 크게 두 가지 부류로 나뉠 수 있다. 첫째는 공공 데이터를 활용한 접근법이다. 해당 데이터에는 인구 통계학적 데이터, 지역별/업종별 업소수와 평균 매출 정보 등이 포함되며, 소상공인시장진흥공단에서 제공하는 상권정보시스템[7]과 서울시가 제공하는 우리마을가게 상권분석서비스[8] 등이 공공 데이터를 이용하여 서비스를 제공하는 대표적인 사례에 해당한다(그림 1).

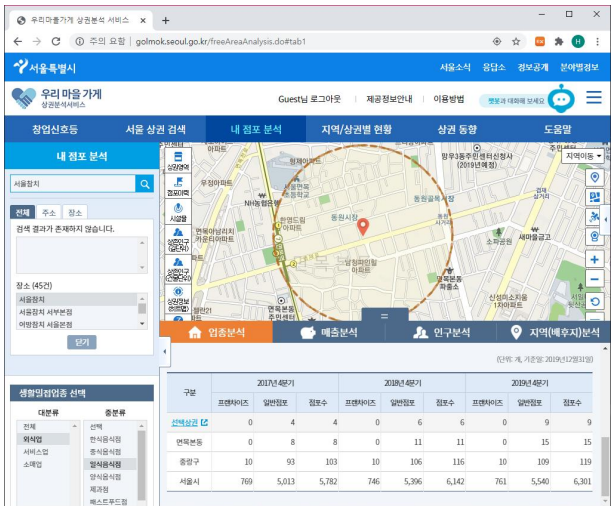

Fig. 1. Business Analysis Service for SMEs (Small and Medium-Sized Enterprises) Provided by Seoul Metro City

[9]에서는 소상공인시장진흥공단이 제공하는 시스템을 이용하여 창업시 수행된 입지 분석 정보와 실제 창업 6개월 이후 나타난 경영성과 간의 상관관계를 분석하였다.

[10]에서는 서울시 상권분석 서비스를 이용하여 잠재 고객을 타겟팅하고 마케팅 활동 수행을 위한 전략들을 도출하였다. 해당 연구에서 볼 수 있듯이, 공공 데이터는 사업 전 입지 분석 등을 수행하는데 효과적으로 활용될 수 있는 반면, 업체에 특화된 데이터가 포함되지 않기 때문에 실제 창업 후의 문제점 및 개선 사항을 도출하는 부분에 있어 한계점이 명확하였다.

둘째는 소셜 빅데이터를 활용하여 현황을 분석하고 마케팅 방향을 도출하기 위한 접근 방법이다. [11]에서는 소셜 빅데이터를 이용하여 전통시장 활성화 요인을 도출하기 위한 연구를 수행하였으며, [12]에서는 소셜 빅데이터와 공공 데이터를 결합하여 지역구 상권의 매출액에 영향을 미치는 요인들을 검증하였다. 이들 연구에서는 소셜 빅데이터의 수집과 키워드 분석을 위해 텍스톰(Textom)[5]을 활용하였으며, 노이즈 리뷰 필터링을 위한 별도의 작업을 수행하였다.

소셜 빅데이터 분석 플랫폼과 관련하여 국내 대표적인 선두 주자는 메조미디어의 티버즈(TIBUZZ)[4], 코난 테크놀로지의 펄스-k[13], 솔트룩스의 블루볼트[14] 등이 있으며, 이들은 주로 일정 규모 이상의 기업들을 대상으로 대규모 SNS 데이터를 활용하여 브랜드 이미지나 경쟁사 비교 분석 서비스를 제공하고 있다. 그러나 소상공인들이 이들 서비스를 이용하기에는 앞서 언급한 바와 같이 비용 측면에서 어려움이 있다. 더불어 이들 플랫폼들은 수집된 리뷰가 풍부하지 않을 때 분석의 정확도에 어떤 영향을 미칠 수 있는지 등과 관련하여 검증이 충분히 이루어지지 않고 있다. 소상공인의 경우 대기업이나 공기업에 비해 대중들에게 덜 알려져 있어 수집되는 리뷰가 현저히 적을 수 있으며, 이 경우 기존 분석 플랫폼을 활용하는데 어려움이 있을 수 있다. (관련 내용은 다음 장에서 자세히 논의한다.)

현재까지 소상공인들이 직접 빅데이터를 수집하고 분석할 수 있는 시스템과 관련해서는 연구가 활발히 진행되지 않고 있다. 소상공인을 대상으로 한 빅데이터 분석 시스템과 관련하여 [15]에서 설계 내용을 제시하고 있으나, 해당 시스템에서는 소셜 빅데이터가 아닌 자체 제작한 스마트 앱을 통해 데이터를 수집하도록 설계되었으며, 이 경우 분석에 필요한 데이터를 수집하고 현장에서 활용되는데 상당한 시간이 소요될 수 있다. 더불어 분석 결과의 정확도를 높이기 위해서는 수집될 데이터의 특성에 대한 파악이 필수적이나, 해당 논문에서는 관련 내용이 제시되지 않고 있다.

III. Characteristics of Big Data Analysis for SMEs

소상공인을 대상으로 한 빅데이터 분석 서비스를 제공하는데 있어 가장 중요한 부분은 앞서 언급한 바와 같이 비용적인 측면이다. 본 연구팀이 ㈜온굿플레이스와 함께 통영시와 거제시 상인들을 대상으로 빅데이터 분석 가격에 대한 수요 조사를 한 결과, 응답자의 90% 이상이 월 20만원 이하의 가격이 적절하다고 응답하였다. 따라서 통신, 금융 등을 포함한 여러 형태의 빅데이터 중 무료로 획득 가능한 소셜 빅데이터가 비용 측면에서 가장 효과적으로 이용될 수 있으며, 본 연구 역시 소셜 빅데이터를 분석에 이용하는 것으로 가정한다.

소셜 빅데이터를 분석할 때 애로 사항 중 하나는 수집된 리뷰(네이버 및 다음 블로그 글, SNS 리뷰 등) 중 주어진 키워드와 연관성이 떨어지는, 이른바 “노이즈” 리뷰를 걸러내야 한다는 점이다. 아래는 본 연구팀이 ㈜온굿플레이스와 함께 상인들에 대한 소셜 빅데이터 분석을 수행하여 파악한 노이즈 리뷰 패턴 중 일부를 소개하고 있다.

○ 다른 상호에 대한 리뷰: 예를 들어 “통영 해물가”를 키워드로 전달할 경우, 수집된 리뷰에 “해물가” 왼쪽/건너편/오른쪽에 위치한 “꿀빵집” 또는 “해물가” 건물 2층에 위치한 “미용실” 등 다른 상호와 관련된 리뷰
○ 홍보성 리뷰: 가게 주인 등이 홍보를 위해 동일 아이디로 다수의 글을 반복적으로 올리는 경우
○ 바이럴 마케팅 리뷰: 글 하나에 여러 식당이나 카페를 해시태그로 함께 언급하던지, 또는 전혀 관계없는 글(노래 가사 등)에 홍보하고자 하는 업체를 해시태그로 언급
○ 기타: 버스정류장 등의 교통 및 지자체 정보 안내에 상호명이 함께 언급되는 경우

아래 표는 2019년 4월 한 달 기준, 통영시와 거제시의 인기 식당 및 카페, 숙박 업소 각각 20곳에 대해 수집된 리뷰 수 대비 노이브 리뷰 수의 비율을 조사한 결과이다. 노이즈 리뷰 판정과 관련하여, 리뷰 내용 중 위에서 언급한 패턴이 있을 경우 노이즈 리뷰로 간주하였다.

Table 1. Ratio of Noise Reviews of popular 20 stores in Tongyoung and Geoje cities

<Tongyoung City>

	Total Reviews	Noise Reviews	Ratio(%)
Restaurant	3,400	1,048	30.8
Cafe	1,946	520	26.7
Hotel	2,190	1,731	79.0
Sum	7,536	3,299	43.8

<Geoje City>

	Total Reviews	Noise Reviews	Ratio(%)
Restaurant	3,200	796	24.8
Cafe	2,900	1,289	44.4
Hotel	2,000	1,423	71.1
Sum	8,100	3,508	43.3

위 표에서 볼 수 있듯이, 전체 리뷰 중 노이즈 리뷰의 비율이 평균 43%에 달하고 있으며, 이들을 포함하여 분석을 수행할 경우 정확도에 심각한 영향을 미칠 수 있다. 특히 숙박업소의 경우 노이즈 비율이 70% 정도에 해당하는데, 그 이유는 숙박업소들이 주로 대형 업소이며 대중들에게 많이 알려져 있어 사람들이 약속 장소를 잡거나 근처 가게들이 자신들을 홍보할 때 해당 숙박업소를 자주 언급하기 때문으로 분석된다.

소상공인을 대상으로 한 소셜 빅데이터 분석에서 파악된 더욱 심각한 문제점은 다수 소상인들의 경우 리뷰가 수집되지 않는다는 점이다. 아래 표는 2019년 4월 한 달 기준, 통영시와 거제시에서 분석을 의뢰한 식당 및 카페, 숙박업소 중 리뷰가 10건 이하인 업체 수와 비율을 조사한 내용이다.

Table 2. Ratio of SMEs whose number of social reviews is less than or equal to 10

<Tongyoung City>

	Stores	Stores whose reviews <= 10	Ratio(%)
Restaurant	62	19	30.6
Cafe	38	6	15.8
Hotel	56	34	60.7
Sum	156	59	37.8

<Geoje City>

	Stores	Stores whose reviews <= 10	Ratio(%)
Restaurant	92	16	17.4
Cafe	64	16	25
Hotel	61	34	55.7
Sum	217	66	30.4

표에서 볼 수 있듯이, 리뷰 수가 10건 이하인 업체의 비율이 평균 35% 정도를 차지하고 있으며, 해당 비율은 분석을 의뢰한 유명 업체가 아닌 일반 업체로 확대할 경우 훨씬 커질 것으로 예상된다. 따라서 상용화가 가능할 정도의 분석 서비스를 제공하려면, 리뷰 수가 적은 경우에도 상인들이 받아들일만한 효과적인 내용을 제시할 수 있어야 한다.

위 문제점에 대한 대안으로, 본 연구에서는 특정 업체의 리뷰가 적을 경우, 지역 내 동종 업계에 포함된 타 업체의 정보를 바탕으로 업계 현황을 분석하여 업종에 대한 가이드 형태로 제시하고자 한다. 예를 들어, 지역 내 업종별로 분석 점수가 높은 상위권 업체들의 현황 정보를 제시하거나, 업종 내 각 분위에 속한 업체들의 정보를 요약하여 가이드 형태로 제시할 수 있다. 이외에도 인구통계학적 정보 등의 공공 데이터를 접목시켜 해당 서비스를 더욱 풍부하게 확장할 수 있다. 관련 내용은 다음 장에서 논의한다.

IV. Service Definition

소상공인 대상 빅데이터 분석 플랫폼을 만들기 위해서는 먼저 소상공인들의 마케팅에 도움이 될 수 있는 서비스 지표(기능)에 대한 정의가 선행되어야 한다. 지원 가능한 마케팅 지표는 어떤 형태의 데이터를 활용할 수 있느냐에 아래와 같이 제공될 수 있다.

1. Public Statistical Data

먼저 통계청 등에서 제공하는 공공 데이터를 이용하여 기존의 상권 분석 서비스가 제공하던 일반적인 마케팅 분석 내용을 제공할 수 있다. 아래의 서비스는 지역별(시군구별)로 제공될 수 있으며, 5장에서 아래 서비스에 대한 구현 화면을 제시한다.

○ 소비인구 진단
- 연령별/성별 인구분포
- (소비 여력이 큰) 30~40대의 인구 비율에 대한 인근 지역과의 비교/분석

○ 지역상권 진단
- 업종별 분포 현황 지도
- 3년간 업종별 업체 수 추이 그래프

2. Social Reviews

다음으로 업체별로 수집되는 소셜 리뷰를 활용하여 마케팅 분석 결과를 제공할 수 있다. 이 경우 앞서 언급한 바

와 같이 유의미한 분석 결과를 제공하기 위해서는 업체별로 최소 10개 이상의 리뷰가 수집 가능해야 한다. 아래는 해당 조건을 만족하는 경우 제공될 수 있는 서비스를 나열하였다. (5장에서 구현 화면을 제시한다.)

○ 사업장 평판 및 리뷰
- 인기지수 및 추이 (아래에 내용 설명)
- 소비자 관점에서 사업장의 장점 및 특징을 알려주는 대표 키워드
- 업체 리뷰 및 이미지

○ 경쟁업체 진단
- 인기 기준 진단 (인기지수가 유사한 업체들과의 비교)
- 업체위치 기준 진단 (인근 업체들과의 비교)
- 업종 기준 진단 (동종 업체들과의 비교)

소상공인 업체들의 분석과 관련하여, 가장 먼저 수행되어야 할 부분으로 해당 업체들의 인기나 인지도(평판)를 정량적으로 나타내기 위한 지표값을 정의해야 한다. 그 이유는 정량적인 지표값이 있어야 업체들을 객관적으로 비교할 수 있기 때문이다. 본 연구에서는 업체 비교를 위한 정량값으로 "인기지수"를 이용하고자 한다. 업체별 인기지수는 온라인에서 수집된 리뷰의 수를 기반으로 아래와 같이 산정될 수 있으며, 온라인 상에서 해당 업체의 평판(인지도) 정도를 정량화한 값에 해당한다.

$$f = \sum_{i=1}^{N} \{n_i w_i - m_i u_i\}$$

위 식은 월별로 수집된 리뷰 수에 가중치를 곱한 값을 합산하고 있다. 보다 자세하게, n_i와 w_i는 i번째 달의 긍정(positive) 리뷰 수와 가중치에 해당하며, 가중치는 가장 최근 달부터 순차적으로 값이 줄어들도록 정의된다. 그리고 m_i와 u_i는 i번째 달의 부정(negative) 리뷰 수와 가중치에 해당하며, 일반적으로 u_i는 w_i에 비해 큰 값으로 정의된다. (부정 리뷰 추출 방법에 대해서는 다음 장에서 설명한다.) 아래 표는 최근 1년(N = 12)동안 수집된 리뷰 수를 바탕으로 인기지수를 산정한 예를 보여준다.

Table 3. Estimation of the popularity score based on the number of reviews collected for one year (N = 12)

	12	11	10	9	8	7	6	5	4	3	2	1	Sum
n_i	27	25	23	17	12	15	13	16	19	22	21	29	
w_i	1.5	1.4	1.3	1.2	1.1	1	1	0.9	0.8	0.7	0.6	0.5	
$n_i w_i$	40.5	35	29.9	20.4	13.2	15	13	14.4	15.2	15.4	12.6	14.5	239.1
m_i	3	2	2	0	1	0	1	0	1	1	2	2	
u_i	3	2.8	2.6	2.4	2.2	2	2	1.8	1.6	1.4	1.2	1	
$m_i u_i$	9	5.6	5.2	0	2.2	0	2	0	1.6	1.4	2.4	2	31.4
Popularity score													207.7

위 식을 통해 업체별로 얻어진 인기지수를 바탕으로, 지역 내 업체들에 대한 정량적인 비교가 가능하다. 예를 들어, 인기지수가 유사한 업체들을 경쟁 업체로 간주하고 위에서 제시한 "경쟁업체 진단" 기능을 수행할 수 있으며, 업종이나 판매 메뉴를 추가적으로 고려한 상세 비교의 구현 역시 가능해진다.

업체별 대표 키워드는 수집된 리뷰로부터 얻어질 수 있으며, 소비자 관점에서 사업장의 장점 및 특징을 알 수 있도록 한다. 키워드 추출을 위해 리뷰별로 형태소 분석이 수행되며, 이를 통해 얻어진 단어 집합으로부터 명사나 형용사를 추출하여 후보 집합을 구성한다. 그리고 후보 집합에서 N회 이상 나타난 단어를 대상으로 빈도순으로 정렬함으로써 최종적으로 업체별 대표 키워드 리스트를 구성할 수 있다.

3. Regional Summary Information

업체별로 얻어진 인지도 점수와 키워드 정보를 지역별로 요약/합산함으로써, 소상공인의 마케팅 경쟁력 향상을 위한 더욱 다양한 정보를 얻을 수 있다.

○ 지역상권 정보
- 지역별 인기 업종
- 지역별 인기 메뉴 및 관련 사업장 리스트
- 업종별 인기지수 추이 및 인기 사업장 리스트

○ 경쟁력 향상을 위한 가이드
- 업종별 인기 사업장 특징
- 업종별 스탠다드 및 분위 현황

먼저 지역별 인기 업종은 지역 내 업종별로 업체들의 평균 인지도 점수를 비교함으로써 추정해 낼 수 있다. 인기 메뉴의 경우 지역 내 업체들의 키워드를 합산하여 메뉴 관련 키워드만 추출하여 출현 빈도수를 비교함으로써 알아낼 수 있으며, 추출된 인기 메뉴를 포함한 업체들을 관련 사업장 리스트에 포함시켜 함께 제시할 수 있다. 업종별 인기 지수 추이 역시 지역 내 업종별로 업체들의 인지도 점수를 월별로 합산하여 구해낼 수 있으며, 최근 인기가 높은 업체들을 위주로 업종별로 인기 사업장 리스트를 구성할 수 있다.

그리고 업종별 인기 사업장 중 인기지수가 최상위인 업체를 추출하여 그들의 키워드와 리뷰를 제공함으로써, 동종 업계에 속한 업체들이 마케팅 경쟁력 향상을 위한 가이드로 참고할 수 있도록 하였다. 또한 업종별로 인지도 점수를 기반으로 업체들을 5분위로 나눌 수 있으며, 자신이 어느 분위에 해당하는지, 다음 분위로 업그레이드하기 위해서는 인기지수를 어느 정도 높여야 하는지 명확한 수치로 제시함으로써 소상공인들의 현황 파악 및 경쟁의식 고

취에 도움을 줄 수 있다.

무엇보다 위에서 제시한 서비스는 소셜 리뷰가 부족한 업체들에게도 적용 가능하다. 따라서 리뷰가 부족한 다수의 소상공인들을 대상으로 해당 서비스를 제공하여 분석 서비스의 만족도를 높이기 위해서는 지역별 요약 정보의 축적이 필수적일 것으로 예측된다.

V. Prototype System

이번 장에서는 위에서 도출된 기능을 구현한 프로토타입 시스템에 대해 소개한다. 프로토타입 시스템은 클라이언트와 서버로 구성되어 있으며, 클라이언트는 HTML5 기반의 모바일 앱(App) 형태로 구현되었다. 그리고 해당 서비스는 KT Cloud 서버[16]에서 동작하도록 구현되었으며, CentOS 7 운영체제와 Tomcat 8.5 웹서버를 기반으로 구현되었다.

1. System Structure

아래 그림은 프로토타입 시스템의 구조를 도식화하고 있다. 프로토타입 시스템은 크게 리뷰 수집기, 형태소 분석기, 노이즈 리뷰 필터, 업체별 요약정보 추출기, 지역 요약정보 추출기, 분석결과 시각화 등의 모듈로 구성되어 있다.

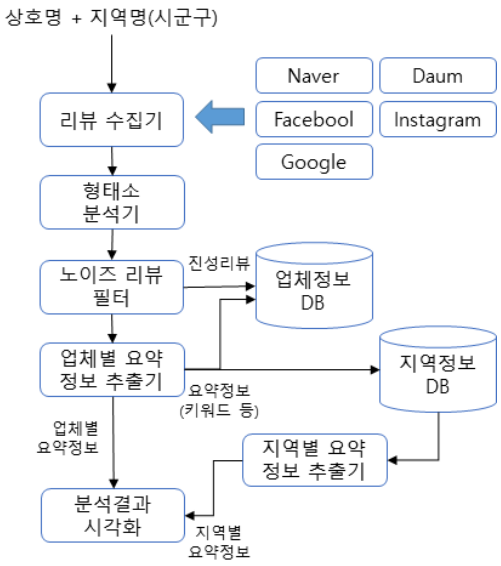

Fig. 2. Architecture of Our Prototype System

시스템의 동작 과정은 다음과 같다. 먼저 분석을 위한 키워드로 업체의 상호명과 지역명(시군구 단위)을 입력 받는다. 예를 들어 통영의 동광식당의 경우 “동광식당 통영”의 형태로 검색 키워드를 구성한다. 다음으로 입력된 키워드에 대해 포털 또는 SNS 업체의 API를 활용하여 리뷰를 수집한다. 현재 데이터 수집을 위해 이용 가능한 API는 네이버, 다음, 구글, 페이스북, 인스타그램 등이다.

데이터 수집이 완료되고 나면, 형태소 분석을 통해 각 리뷰에 대한 단어 집합을 추출하며, 구현에는 KAIST에서 제작한 한나눔 형태소 분석기[17]가 이용되었다. 추출된 단어 집합과 리뷰를 기반으로 노이즈 리뷰 필터링 과정이 수행되며, 현재 구현에는 간단한 패턴 매칭을 이용하여 기본적인 필터링만 수행되어 정확도가 낮은 문제점이 있다.

필터링 이후 남은 진성 리뷰들은 업체별 요약정보 추출기를 통해 인기지수 산정 및 키워드 추출 등의 작업이 이루어지며, 추출된 요약정보는 업체정보 DB와 지역정보 DB에 함께 반영된다. 특히 인기지수 산정을 위해 진성 리뷰들에 대한 긍정/부정 리뷰 여부를 구분하는 작업이 수행되며, 현재 구현에서는 “맛없는”, “불친절한” 등의 미리 정의된 패턴이 나타날 경우 해당 리뷰를 부정으로 간주한다. 부정 리뷰 추출 역시 기본적인 패턴 매칭이 이용되므로 정확도 개선을 위한 추가 연구가 필요한 상황이다.

이후 지역별 요약정보 추출기에서 업데이트된 지역정보를 얻어와 업체별 요약정보와 함께 분석결과 시각화 모듈로 전달한다. 시각화 모듈은 전달받은 분석 결과를 모바일 디스플레이 및 PC 화면에 맞게 최적화하여 결과를 보여준다.

2. Visualization of Analysis Results

아래 그림은 프로토타입 시스템에서 분석 결과를 시각화한 예이다.

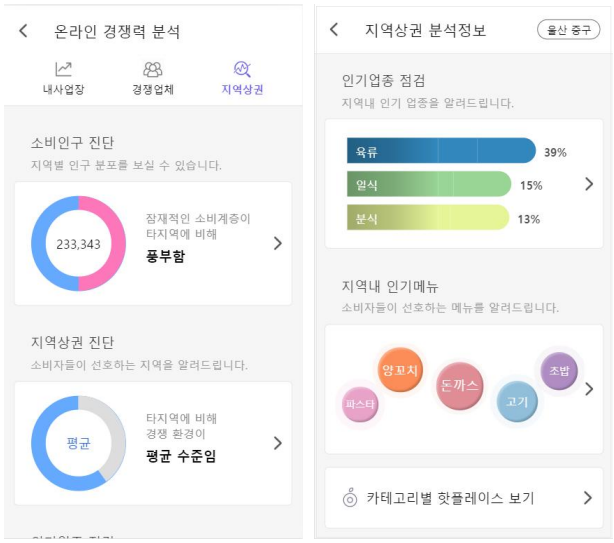

Fig. 3. Regional Commercial Analysis Information for Ulsan Jungu Provided By Our Prototype System

[그림 3]은 울산광역시 중구를 대상으로 지역상권 분석을 수행한 결과를 시각화한 화면이다. 왼쪽은 공공데이터를 기반으로 얻어진 소비인구 및 지역상권 진단 결과를 보여주며, 오른쪽은 지역별 요약정보를 바탕으로 얻어진 지역 내 인기업종 및 인기메뉴 정보를 제시하고 있다.

[그림 4]는 울산 중구의 "제이키친"을 대상으로 소셜 리뷰 분석을 수행한 결과이다. 왼쪽은 업체의 인기지수와 대표 키워드 분석을 수행한 결과이며, 오른쪽은 인기지수 및 업체의 위치를 기준으로 경쟁업체를 진단한 결과이다.

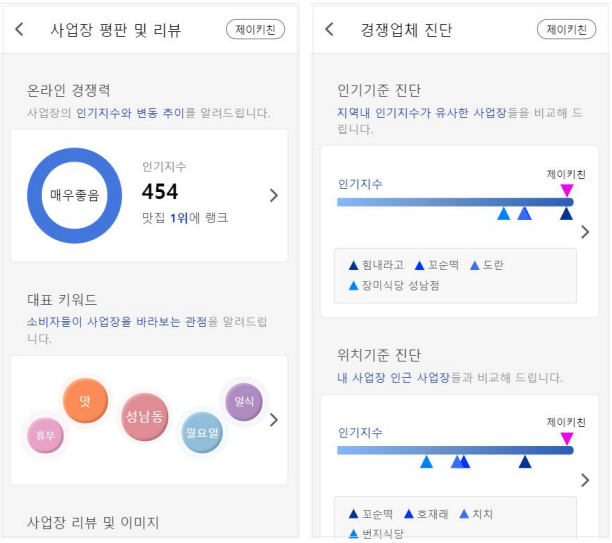

Fig. 4. Social Review Analysis Information for Ulsan Jungu Provided By Our Prototype System

3. Technical Issues

프로토타입 시스템의 구현을 통해 얻어진 기술적 이슈는 크게 세 가지로 요약될 수 있다. 첫째는 형태소 분석기의 분석 효율이 매우 중요하다는 점이다. 예를 들어 현재 프로토타입에서 이용하고 있는 한나눔 형태소 분석기의 경우 리뷰별로 100개 정도의 단어로 구성된 35개의 리뷰를 처리하는데 (Intel Core-i5 기준) 평균 1초 정도의 시간이 소요된다. 만약 전국에 10만개의 인기 업체가 등록되어 있고 업체별 평균 1,000개 정도의 리뷰를 처리해야 한다고 가정할 경우, 리뷰 전체를 분석하는데만 약 33일이 소요된다. 이에 노이즈 리뷰 필터링 및 요약정보 추출 시간까지 합칠 경우, 매월 배치 형태의 일괄적인 리뷰 수집 및 (4.3절에서 제시한) 요약정보 업데이트는 사실상 불가능한 상황이 된다. 이 경우, 지역별 요약 정보에 기반한 마케팅 가이드 및 업종별 분위 정보의 신뢰도에 문제가 생길 수 있어 리뷰 수가 적은 업체들이 분석 서비스를 활용할 때 만족도가 크게 저하될 수 있다. 따라서 분석 정보의 신뢰도를 높이기 위해서는 형태소 분석의 효율성 개선이 필수적이다.

두 번째로 파악된 이슈는 노이즈 리뷰 필터링 관련 내용이다. 3장에서 언급한 4가지 노이즈 리뷰 패턴 이외에도 더욱 다양한 패턴이 있을 수 있으며, 각 패턴마다 서로 다른 필터링 기법이 적용될 수 있다. 따라서 필터링 정확도를 높이기 위해 최대한 많은 노이즈 리뷰 패턴을 찾아내는 것이 중요하며, 각각의 리뷰 패턴을 걸러내기 위한 학습 데이터 및 최적화된 기계학습 기법을 찾아내기 위해 지속적인 노력이 필요할 것으로 보인다.

마지막 이슈는 부정 리뷰 검출과 관련된 내용으로, 단순 패턴 매칭을 이용하는 방법에는 한계가 있을 수 있다는 점이다. 앞서 언급한 바와 같이, 프로토타입 시스템에서는 "맛없는" 등의 단어가 리뷰에 포함되면 부정 리뷰로 간주한다. 그러나 해당 방법은 단어 전후의 문맥 컨텍스트를 파악하지 못해 정확도가 떨어지는 단점이 있다. 예를 들어, "친구는 맛없다고 했는데, 내가 직접 먹어보니 맛있었다." 등의 리뷰는 단순 패턴 매칭 방법으로는 알아내기 어렵다. 이와 관련하여, [17-19] 등에서 소개된 감성 분석 또는 기계학습 기반의 기존 연구를 도입하여 부정 리뷰 검출의 정확도를 개선할 수 있을 것으로 예측된다.

VI. Conclusion and Future Work

본 논문에서는 소상공인을 대상으로 한 빅데이터 분석 시스템을 개발하기 위해 필요한 내용들을 소개하였다. 먼저 소상공인을 대상으로 한 빅데이터 분석이 일반적인 경우와 어떤 측면에서 차이가 있는지 소개하였다. 이와 관련하여 소상공인을 대상으로 분석을 수행할 경우 가격적인 측면으로부터 소셜 빅데이터를 분석에 활용할 것을 제안하였으며, 다수의 영세업체들의 경우 수집되는 리뷰가 부족한 점으로부터 지역별/업종별 요약 데이터를 기반으로 마케팅 경쟁력을 향상시킬 수 있는 가이드 정보를 제공할 것을 제안하였다.

다음으로 업체별로 수집되는 리뷰 데이터와 지역별/업종별로 유지되는 요약정보, 그리고 지역별 공공 데이터 등을 기반으로, 소상공인들의 마케팅에 도움이 될 수 있는 서비스 지표들을 도출하였다. 해당 지표에는 소비인구 진단, 지역상권 진단, 사업장 평판 및 리뷰, 경쟁업체 진단 등이 포함되며, 수집되는 리뷰가 적을 경우에도 효과적으로 이용될 수 있는 마케팅 가이드 정보를 함께 제시하였다.

마지막으로, 도출된 서비스를 기반으로 프로토타입 시스템을 구현하였으며, 시스템 구현에 어떤 기술적인 이슈들이 있는지 파악하였다. 파악된 이슈는 크게 형태소 분석의

효율성과 노이즈 리뷰 필터링의 정확도 및 부정 리뷰 검출 정확도와 관련된 내용이며, 시스템의 완성도를 높이기 위해 해당 부분에 대한 연구를 지속적으로 진행할 예정이다.

본 연구의 한계점으로는 현재 도출된 분석 결과가 경영 분야에서 다루는 마케팅 방법 및 전략 수행 단계별로 필요한 정보 및 활용 방법 등과 충분히 연계되지 못하고 있다는 점이다. 이와 관련하여 소상공인들의 실질적인 마케팅 활동에 도움을 주기 위해서는 마케팅 관점에서 필요한 변수들을 도출하고, 이를 소셜 리뷰를 포함한 다양한 형태(카드 매출, 고객들의 동선을 포함한 기지국 정보 등)의 빅데이터 분석을 통해 얻어내기 위한 심층적인 연구가 필요할 것으로 예측된다.

ACKNOWLEDGEMENT

This research was supported by the Sahmyook University Research Fund in 2020.

REFERENCES

[1] W. L. Kang, H. G. Kim, and Y, J. Lee, "Reducing IO Cost in OLAP Query Processing with MapReduce," IEICE Trans. Inf. & Syst, Vol. E98-D, No. 2, pp. 444-447, Feb. 2015.

[2] K. H. Lee et al., "Parallel Data Processing with MapReduce: a Survey," ACM SIGMOD Record, Vol. 40, No. 4, pp. 11-20, 2012.

[3] IDC Korea, https://www.idc.com/getdoc.jsp?containerId=prAP 45938720

[4] MezzoMedia Tibuzz, http://tibuzz.co.kr/main/about

[5] Textom, http://www.textom.co.kr/home/main/main.php

[6] OnGoodPlace Inc., http://www.ongoodplace.com

[7] Small Business Promotion Agency, http://sg.sbiz.or.kr/

[8] Seoul Metro City, https://golmok.seoul.go.kr/main.do

[9] Y. S. Eu, "A study on the Relationship between the Commercial District Information System and Franchisees Management Performance," FoodService Industry Journal, Vol. 14, No. 1, pp. 95-101, Mar. 2019.

[10] M. H. Cho et al., "Customized Marketing Strategy Support Services for Small Business," Proceedings of the Spring Conference of Korean Institute of Industrial Engineers, pp. 2618-2621, 2016.

[11] S. H. Park, and H. C. Lee, "The Traditional Market Activation Factor Derivation Research Through Social Big Data," Seoul City Research, Vol. 19, No. 3, Sept. 2018.

[12] I. J. Yoo, B. G. Seo, and D. H. Park, "Smart Store in Smart City: The Development of Smart Trace Area Analysis System Based on Consumer Sentiments," Journal of Intelligent Information Systems, Vol. 24, No. 1, pp. 25-52, Mar. 2018.

[13] Konan Tech. Pulse-K, https://www.pulsek.com/

[14] Saltlux BlueBolt, http://www.saltlux.com/bigdata/bluebolt.do

[15] J. O. Song, J. H. Cho, and S. M. Lee, "Design and Implementation of Marketing System for Traditional Markets based on Big-Data," Proceedings of the Winter Conference of KSCI, pp. 191-192, Jan. 2018.

[16] KT Cloud Server, http://cloud.kt.com

[17] P. S. Jang, "Study on Principal Sentiment Analysis of Social Data," Journal of the Korea Society of Computer and Information, Vol. 19, No. 12, pp. 49-56, Dec. 2014.

[18] B. H. Baek, I. K. Ha, and B. C. Ahn, "An Extraction Method of Sentiment Infromation from Unstructed Big Data on SNS," Journal of Korea Multimedia Society, Vol. 17, No. 6, pp. 671-680, June 2014.

[19] Mary M. Flory, "How are You Feeling? Sentiment Analysis Aims to Find Out," AMA Access: Marketing Researchers, Apr. 2011.

Authors

Hyeon Gyu Kim received the B.S. and M.S. degrees in Computer Science from University of Ulsan, and Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology, Korea, in 1997, 2000 and 2010, respectively. Dr. Kim joined the faculty of the Division of Computer Science and Engineering at Sahmyook University, Seoul, Korea, in 2012. He is currently an Associate Professor in the Division of Computer Science and Engineering, Sahmyook University. He is interested in big data processing, data stream processing, and mobile computing.