<가설 설정>

1. 빅 데이터의 유행과 확산은 사람들에게 이전의 방식보다 더 큰 이점을 준다는 인식과 언론에서의 노출을 통해 자신에게 적합하다는 판단이 영향을 주었다고 가정한다. (Factorial ANOVA)

2. 빅 데이터의 유행과 확산은 오피니언 리더들의 빅 데이터에 대한 언급과 주장의 증가로 인해 사람들이 빠르게 이 기술에 대해서 받아들이고 있다.(F-test)

3. 과거와는 다른 지식 창출, 공유, 습득의 기술이 빅 데이터에 관한 관심에 영향을 줬을 것이다.(F-test)

4. 빅 데이터를 활용한 트렌드 분석과 마케팅 활용은 빅 데이터에 관한 관심에 영향을 주었을 것이다.(F-test)

5. 비즈니스, 행정, 정치, 외교, 안보 등 여러분야에 걸친 빅 데이터 분석의 중요도가 빅 데이터 관심도 증가에 영향을 미칠 것이다.(Regression)

6. 언론에 빅 데이터 라는 키워드가 노출되는 빈도수가 높을수록 더많은 관심을 가질 것이다. (Regression)

7. 빅 데이터를 이용한 기업의 매출과 그렇지 않은 기업의 매출의 차가 빅 데이터에 관한 관심에 영향을 주었을 것이다.(t-test)

8. 빅데이터 솔루션을 적용한 회사의 매출과 저장된 데이터의 양, 그 회사가 사람들에게 노출된 정도 등이 빅 데이터에 대한 관심에 영향을 끼칠 것이다.(multiple regression)

<가설 종류 및 정리>

IV

1. 언론의 노출횟수 (숫자)

2. 지식 창출, 공유, 습득 방식에 대한 변화 (사회 현상학적 연구자료 검색)

3. 트렌드 분석과 마케팅 활용 (실제 사용되는 기술 발전 추이와 사람들의 인식 설문조사)

4. 빅 데이터라는 키워드 노출 빈도 확인(실제 노출되는 데이터를 추출)

5. 기업의 빅 데이터를 이용 유무의 차이로 인한 매출 차이(데이터 분석을 하는 기업과 아닌 기업의 매출차이)

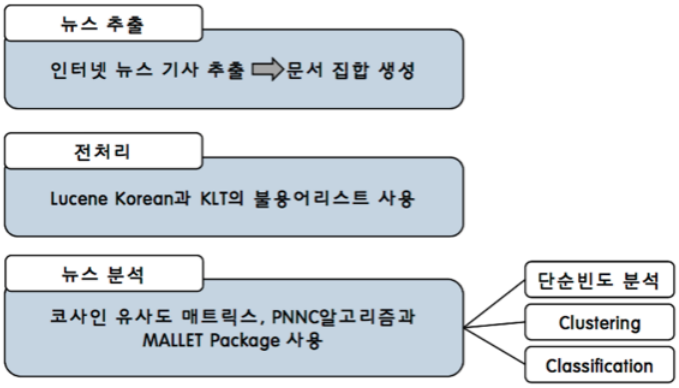
6. 오피니언 리더들의 언급과 주장 (사람들의 설문조사 & 실제 언급에 대한 횟수)

CV : 기업의 크기(중소기업) , 사람들 인식 (20-30대를 중심으로 실제 온라인 이용자)

DV : 빅 데이터의 유행과 확산

<데이터 추출 방법 및 문제시 차선책 방안>

1. 언론의 노출횟수 (숫자) : 언론의 노출횟수의 경우, 실제로 온라인 상에서 얼마만큼 빅 데이터가 언급 되었는지 언론사의 데이터를 가지고 와서 빅 데이터 관련 단어를 중심으로 얼마나 언급된 수가 늘어났는지 확인한다.



이와 같은 방식으로 뉴스를 추출하여 전처리 과정을 통해서 단어를 솎아낸 후에 뉴스를 분석한다. 이 그림의 과정에서는 다양한 알고리즘을 사용했지만 우리가 원하는 것은 이것이 긍정적인지, 부정적인지를 확인하는 것이 아니라 언급한 횟수에 대한 부분이기 때문에 그 횟수의 증가 추이를 확인하기 위해서 빈도수 검사를 실행한다.

(그림1)

2. 지식 창출, 공유, 습득 방식에 대한 변화 (사회 현상학적 연구자료 검색) : 사회 현상학적인 연구자료를 확인한다. 예를 들어, 빅 데이터를 실제로 이용하는 분야가 어떻게 증가 하였는지, 그 증가에 대한 부분이 어떻게 발전하게 되었는지 등 실제 사회에서 빅 데이터에 관심을 가질 수 밖에 없는 현상학적인 이유를 찾아본다. 예를 들어, '컴퓨터의 발전과 인터넷의 발전으로 데이터 베이스를 구축하는 서버가 발전하고, 그 서버를 많은 사람들이 이용하는 검색 엔진에서 이용하게 되고 그러한 매스 데이터가 쌓이게 되면서 자연스럽게 발전했을 것이다'라는 식의 컴퓨터 발전과 사회 발전의 관계를 분석한 논문 등을 참고하여 여러가지 사회 현상학적인 요인들을 결합하여 결과를 얻는다.

3. 트렌드 분석과 마케팅 활용 (실제 사용되는 기술 발전 추이와 사람들의 인식 설문조사) : 실제 기술이 발전된 추이를 보여주는 데이터를 보여주고, 사람들의 인식의 설문조사와 상관관계를 보여 준다. 이 경우, 년도 별 사람들의 설문조사를 얻을 수 없을 것이다. 예를 들어, 2005년 사람들의 빅 데이터에 대한 인식을 확인하는 데이터는 얻을 수 없을 것이다.

-> 따라서 이러한 경우는 이전에 빅 데이터라는 분야에 대해서 언급이 된 시점을 조사하여 그 당시 사람들이 빅 데이터에 대한 논문(2000년대)을 찾아보고, 어느정도의 예측으로 대신한다. 예를 들어, 아래 그림(2015년도 조사 자료)과 같이 업종별 미도입 사유에 대한 자료가 있다. 이 데이터가 현재(2016년)변화가 있다면 사람들이 혁신이라고 생각하는 빅 데이터의 상대적 이점, 적합성, 복잡성, 시험 가능성 등 다양한 요인의 변화가 있거나 혁신이 확산되는 단계가 변화 되었다고 볼 수 있다고 생각한다. 이처럼 기존의 다양한 데이터를 중심으로 기술 발전 추이와 함께 비교한다면 불가능한 설문조사나 부족한 질적조사를 대체할 수 있다고 생각한다.

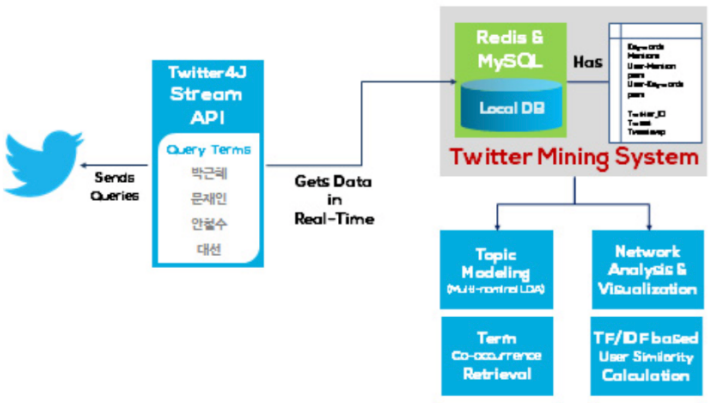
[표 8] 업종별 빅데이터 미도입 사유 [중복 응답, 단위 : 응답기업수]

구분	도입 관심 수준						
	빅데이터라고 할 만한 데이터가 없음	빅데이터 도입 효과에 대한 불신	관련 전문 인력 없음	CEO/CIO 무관심	빅데이터를 분석할 만큼 큰 기업이 아님	빅데이터 자체가 어떤 것인지 잘 모름	빅데이터 도입효과가 나타날 업무가 없음
공공	59	16	20	23	29	13	23
금융	6	4	2	6	3	2	3
유통/서비스	60	18	18	10	41	11	19
제조	119	26	26	60	74	36	59
의료	34	21	9	10	20	8	12
통신/미디어	11	7	3	3	9	2	6
Total	289	92	78	112	176	72	122

규모별로 살펴볼 경우 기업 규모가 작을수록 자사가 보유한 데이터양의 문제로 빅데이터 도입을 꺼리는 기업이 많았으며, 전문 인력의 문제 및 CEO/CIO의 무관심이 큰 것으로 나타났다.

(그림2)

4. 빅 데이터라는 키워드 노출 빈도 확인(실제 노출되는 데이터를 추출) : 빅 데이터가 얼마나 자주 노출 되는지 확인하기 위해서 트위터 API를 이용해서 데이터를 추출한다. 트위터에서 데이터를 추출하는 이유는 트위터가 사람들이 가장 많이 실시간으로 자신의 생각을 쓰고, 텍스트로 되어 있는 자료이기 때문에 얼마나 많이 나오는지 확인하는 추이를 확인하기 용이하다고 생각했기 때문이다.



트위터API 사용 방식은 우선 트위터에서 데이터를 긁어서 가지고 온 후에 트위터 마이닝을 해주는 프로그램을 선정(자바, R 등)한 후에 그에 맞도록 데이터를 처리하여 빈도수 검사를 하여 사용한다. 실제로 트위터에서 데이터를 추출하여 검색하는 방식은 많이 사용되고 있기 때문에 추출하여 빈도를 찾아보는 빈도수 검사의 경우 어려움이 없을 것이라고 생각한다.

(그림3)

5. 기업의 빅 데이터를 이용 유무의 차이로 인한 매출 차이(데이터 분석을 하는 기업과 아닌 기업의 매출차이): 우선 데이터를 실제로 얼마만큼 보유하고 있는지 그 추이를 확인하고, 그것을 분석할 수 있는 능력의 차이를 조사한다. 그 조사를 통해서 기업의 매출의 차이가 그러한 데이터의 차이와 가공 기술 및 활용 능력의 차이로 인해 매출이 차이가 난다는 것을 보여 준다. 이러한 방법으로 우선 기존의 기업을 분석한 다양한 자료를 찾아본다. 이러한 자료는 우선 비교 할 기업을 선정해야 하는데, 실제로 빅 데이터를 이용하여 큰 매출을 이어진 회사는 IT 분야와 마케팅 분야가 많을 것이라고 생각하여 그러한 기업을 중심으로 선정한다. 앞서 이야기 한 것 처럼 대기업이 아닌 중소기업으로 통제하여 조사한다.

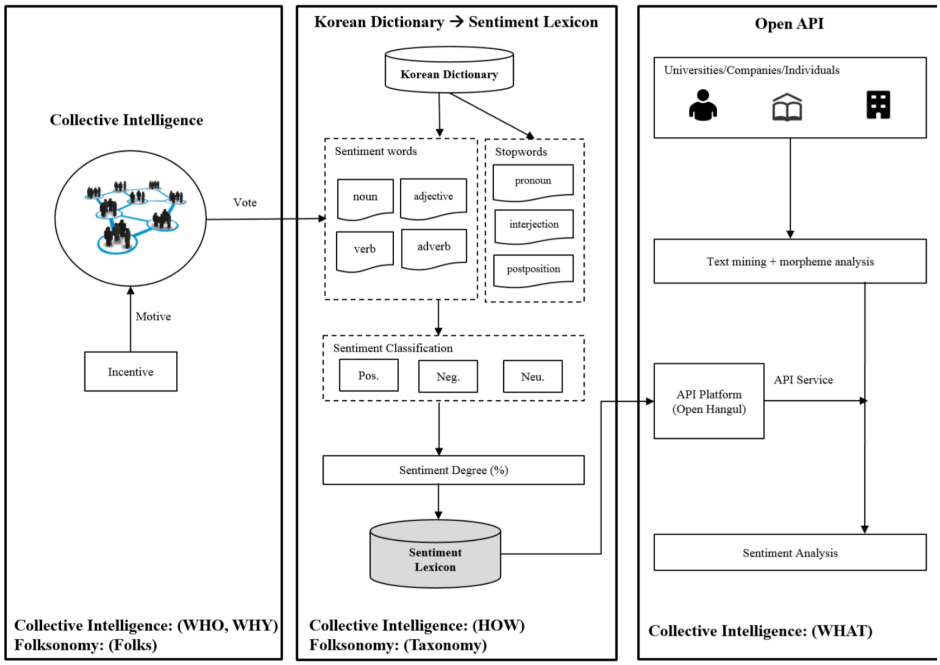
(그림4)

다음 그림은 2011년 미국 산업별 평균 데이터 보유량이다. 이와 같이 산업 별 데이터를 얼마만큼 보유 하고 있는지를 시대별로 정리하고 가장 많은 분야가 가장 데이터를 활용한다는 근거가 있다면 그 분야의 매출이 빠르게 증대한다는 것을 보여주는 방법도 있다고 생각한다. (그림 4).

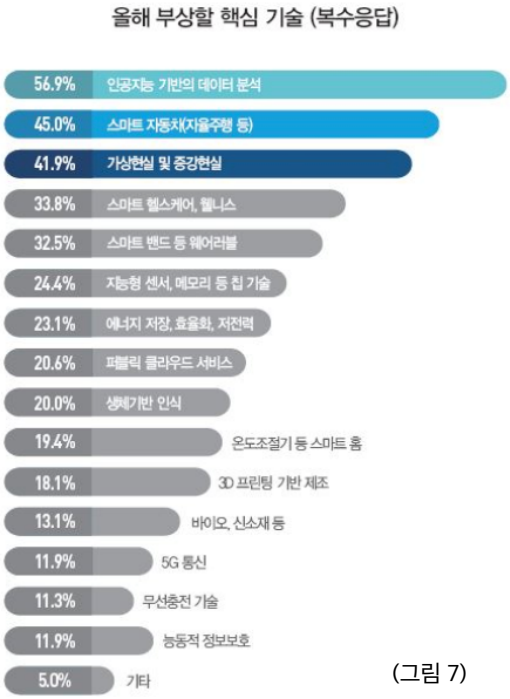
(그림 5)

또한 실제로 적용되어야 하는 분야의 종사자의 설문조사 자료를 통해서 어떻게 변화하고 있는지를 조사한다면 양적인 데이터를 중심으로 하지만 질적인 부분까지도 어느정도 충족할 수 있다고 생각한다. 하지만 이러한 데이터의 경우 없는 경우가 있을 수 있기 때문에 참고자료 정도로 이용하는 편이 좋을 것이라고 판단한다.

6. 오피니언 리더들의 언급과 주장 (사람들의 설문조사 & 실제 언급에 대한 횟수): 실제로 오피니언 리더들이 얼마나 언급 했는지 데이터를 얻기 위해서 가장 먼저 트위터의 오피니언 리더를 중심으로 확인한다. 그 방식으로 집단 지성과 '빅 데이터'라는 단어가 얼마나 연결되어 이루어지고 있는지를 확인한다. 또한 빅 데이터 관련 서적과 논문이 얼마나 많이 출판되고 있는지 조사한다. 또한 그 판매량이 얼만큼인지를 확인한다. 이러한 자료를 바탕으로 전문가들이 얼마나 많이 빅데이터를 언급하고 있는지 확인할 수 있고, 사람들이 얼마나 관심을 갖고 있는지를 서적의 판매량을 통해서 확인해 본다. 아래의 그림은 트위터에서 집단 지성과 단어를 어떻게 관계 하는지 보여주는 과정이다.(그림 6)



(그림 6)



(그림 7)

이 외에도, 언론에서 빅 데이터를 정의하는 다양한 자료가 얼마나 많이 나오는지 확인해 본다. 의제설정이론(Agenda Setting)에 근거하여 이야기 하면 미디어의 의제설정이 실제로 공론화 되는 경우가 많은데, 빅 데이터의 경우에도 그런 의제 중 하나일 수도 있기 때문이다. 미디어가 새로운 기술 중 빅 데이터에 주목을 하고, 그것에 대해서 많이 언급함과 동시에 사람들에게 특정한 이미지를 부여했다면 그것 또한 하나의 가설 설정을 하여 확인해 볼수 있다고 생각한다. 하지만 이것 자체를 하나의 가설로 설정하여 주목하기 보다는 다중회귀분석의 하나의 요인으로 사용하면 좋을 것이라고 사료된다. (그림 7)