빅데이터 활용을 지원하는 수집 데이터의 가공과 정제

김은석 ㈜ 지디에스 컨설팅그룹 대표이사

1. 머리말

1.1 4차 산업혁명의 출발점은?

흔히들 요즘은 4차 산업혁명의 시기라고 한다. 디지털 기술의 발전으로 기존과 차별되는 기술 혁신이 로봇공학, 나노기술, 양자컴퓨팅, 생명공학, IoT, 3D 인쇄 및 자율차량과 같은 기술 전 분야에 확산되고 있다.

4차 산업혁명의 기본 인프라의 근간에는 데이터, 네트워크 및 인공지능이 있다. 근래에는 이를 통칭하여 'DNA(Data-Network-AI)'라고 부르곤 한다. DNA에는 4차 산업혁명을 순조롭게 진행하려면 다양한 데이터를 빠른 네트워크로 수집하여 인공지능을 잘 적용해야 한다는 의미가 담겨 있다. 결국 4차 산업혁명은 데이터에서 시작되며, 그러하기에 데이터가 4차 산업혁명의 '쌀'이라고 불리곤 하는 것이다.

이에 따라 최근 정부기관을 비롯한 각 기관과 기업에서 데이터를 적극적으로 활용하고자 데이터 공개와 판매가 활발히 이루어지고 있다. 그러나 수많은 데이터가 개방되어 거래되고 있음에도 정작 데이터 이용자는 쓸 만한 데이터가 부족하다고 아우성이다. 본고에서는 '데이터 가공 및 정제'에 대한 고찰을 통해 이러한 '데이터 공급과 수요 간 불일치'의 원인을 찾아보고자 한다.

1.2 데이터 가공과 정제란?

데이터를 잘 활용하려면 데이터 수집, 가공 및 분석, 활용의 전 과정이 유기적으로 연계되어야 한다. 여기서 데이터 가공 및 정제란 수집된 데이터를 정리하고 표준화하며 통합하는 일련의 과정을 뜻한다. 데이터를 분석하기 전, 분석에 적합한 데이터를 만드는 사전처리 전반을 일컫는다고 할 수 있다.

데이터 분석을 요리 과정에 비유해보자. 이 경우 데이터는 요리의 재료에 해당한다. 음식을 만들려면 우선 재료를 조리과정에 맞게 손질해야 한다. 이 과정이 바로 데이터 가공 및 정제다. 요리를 해본 사람이라면 잘 알겠지만, 요리 재료를 다듬는 일은 귀찮기도 하고 손도 많이 가는 작업이다. 상점이나 마트에서 가공된 재료가 아니라 원재료 자체를 손질해야 한다면 번거로움은 더할 것이다. 만약 쇠고기 스테이크를 만들려면 직접 소를 잡고, 부위별로 나누어 발골해야 한다면 대부분은 포기하고 말 것이다. 결국 재료를 손질하는 일은 고도의 전문성과 노고가 필요한 작업이며, 이 과정이 없으면 제대로 된 요리가 탄생할 수 없는 것이다.

다시 데이터 분석과 요리의 비유로 돌아가 보자. 쇠고기 스테이크라는 맛있는 음식, 즉 데이터

를 활용한 산출물을 만들려면 '좋은 소를 선별(원천 데이터 확보)'하여 '도축장에서 소를 도축하고, 도축한 소의 뼈를 발라내어 부위별로 나누는 작업(데이터 가공)'이 필수적으로 선행되어야 한다. 데이터 분석에서도 의미 있는 분석 결과를 도출하려면 데이터에 내재한 여러 오류를 먼저 찾고(=정제), 같은 형태로 통일(=표준화)하는 작업을 잘 수행하여야 한다.

1.3 인공지능과 데이터 가공 및 정제

한국 정부는 인공지능과 빅데이터를 기존 산업에 결합하여 육성한 스마트 산업을 통해 새로운 성장을 도모하고 있다[1]. 이를 위해 2019년 12월 '인공지능(AI) 국가전략'을 발표하고 인공지능 강국으로 도약하는 것을 목표로 비전과 실행과제를 제시했다. 특히 세계를 선도하는 인공지능 생태계를 조기에 구축하고자 공공기관이 보유한 공공 데이터를 전면 개방하는 한편, 자율주행, 스마트시티 등의 신산업 분야에서 AI 활용을 활성화하기 위해 공공 데이터를 적극적으로 발굴하고 활용하는 데 정책적 노력을 기울이고 있다.

이러한 정책적 노력으로 이전보다 훨씬 많은 양의 데이터가 수집, 생산됐으며, 이전에는 데이터 활용 가치를 인정받지 못한 사진, 동영상, 음성과 같은 비정형 데이터(unstructured data)의 증가가 특히 두드러졌다.

비정형 데이터란 비구조화 데이터, 즉 미리 정의된 데이터 모델이 없거나 정형화된 방식으로 정리되지 않은 정보를 말한다. 따라서 이를 정보로 활용하려면 다양한 사전 처리(preprocessing)가 필요하다. 예를 들어 자동차 번호판을 인식하려면 자동차를 찍은 사진에서 번호판 부분만을 따로 잘라낸다든지, 얼굴을 인식할 때 얼굴을 가린 모자나 머플러는 삭제하고 얼굴 부분만 추출해야 원하는 기능을 제대로 수행할 수 있다.

이처럼 사진, 동영상, 음성 등에서 우리가 관심을 가지고 분석할 대상을 추출하는 작업을 어노테이션(annotation)이라고 하며, 추출된 정보를 효율적으로 분류하기 위해 주석을 부여하는 작업을 라벨링(labelling), 혹은 주석화(註釋化)라고 한다. [그림 1]은 여러 사진에서 분석에 필요한 부분만을 추출하는 작업의 예시이다. 인공지능 자체는 사람의 개입 없이 컴퓨터에 의해 구현되어야 하지만, 자동화된 인공지능의 알고리즘을 구축하려면 많은 양의 비정형 데이터를 사람이 직접 어노테이션하고 라벨링해야 한다.

※출처: SNS 이미지 정보 인공지능 알고리즘 구축 보고서, GDS컨설팅 그룹, 2020

[그림 1] SNS 이미지 정보에서 여행, 레저 이미지 정보 어노테이션

2020년 하반기 한국 정부는 '디지털 뉴딜'의 대표 과제인 '데이터 댐' 사업을 본격적으로 시작했다. 데이터 댐을 구성하는 7대 핵심 사업 중 가장 중요도와 비중이 높은 것이 단연 'AI 학습용 데이터 구축사업'이다. 총 2만 2천 명의 고용이 창출된다고 하는데, 이들 대부분이 비정형 데이터 가공 및 정제에 투입된다.

2. 데이터 활용과 데이터 가공 및 정제

2.1 데이터 가공이 필요한 이유

- 쓸 만한 데이터가 부족하다?

한국의 공공 데이터 개방 건수는 세계 1위이며, 통신이나 카드 등 활용성이 높은 데이터에 대한 거래도 다른 나라에 비해 활발하다. 그러나 정작 데이터를 이용하여 여러 분석이나 서비스를 수행하는 이용자층에서는 쓸 만한 데이터가 없다는 지적이 이어지고 있다.

왜 데이터는 많은데 쓸 만한 데이터는 부족한가? 일각에서는 수많은 데이터 중 유 용한 것이 부족하기 때문이라고 하고, 다른 한편에서는 데이터 활용에 대한 법률적 규제가 심하기 때문이라고 한다. 유용한 데이터가 부족하다고 주장하는 쪽에서는 필요한 데이터의 가짓수를 늘리는 것을, 법률적 규제를 지적하는 쪽에서는 정부가 다른 나라처럼 규제를 더 푸는 것을 해결방안으로 제시한다. 이처럼 여러 의견이 있지만 쓸 만한 데이터가 부족한 문제의 원인을 제대로 파악하려면 실제 현장에서는 데이터가 어떻게 활용되는지 이해해야 한다.

- 데이터 수집 목적과 활용이 괴리되는 이유는?

데이터는 특정 목적을 위해 수집되므로 목적에서 벗어난 용도로 사용하는 경우 종종 '오류 아닌 오류'가 발생한다. 예를 들어 스마트폰 사용 빈도가 성별에 따라 어떻게 다른지 알기 위해 통신회사의 통화 건수를 남녀별로 비교한다고 하자. 통신회사에서는 휴대전화 사용요금을 부과하기 위해 특정 전화를 누가 사용하는지, 얼마나 사용했는지에 대한 기록을 데이터로 축적하여 관리하고 있으므로 이를 분석하면 쉽게 해결 될 문제라고 생각할 수 있다.

그런데 이러한 분석과정에는 과연 아무런 문제가 없을까? 실제 데이터를 분석해 보면 분석목적에 어긋나는 데이터 수집상 문제가 다수 발생한다. 즉, 분석자는 휴대전화 사용자의 성별을 알기 위해 통화기록과 등록자의 주민등록상 성별 데이터를 수집하여 분석을 수행할 것이다. 여기서 발생할 수 있는 오류는 휴대전화의 실제 사용자와 휴대전화 등록자가 반드시 일치하지 않는다는 점이다. 만약 내 명의로 발급받은 전화를 배우자나 다른 가족이 사용한다면 등록자와 사용자의 성별(gender)이 일치하지 않는다. 이 경우 원래의 의도와 다른 분석 결과가 나올 것이며, 이는 명백한 분석 오류다.

그렇다면 통신회사가 데이터를 수집할 때 문제가 있었을까? 통신회사에서 수집하는 이용자정보의 궁극적인 목적은 사용량에 따라 정확한 요금을 부과하는 것이다. 설사 남자가 등록한 전화를 여자가 사용하더라도 사용 내역에 대한 요금만 정확하게 부과됐다면 기업의 영리활동에는아무 문제가 없다.

위의 사례는 데이터의 본래 수집 목적에서 벗어난 용도로 사용할 때 발생하는 오류를 잘 보여준다. 성별에 따른 이동통신 이용 현황은 기업이 요금을 부과하는 것과는 목적이나 용도가 다

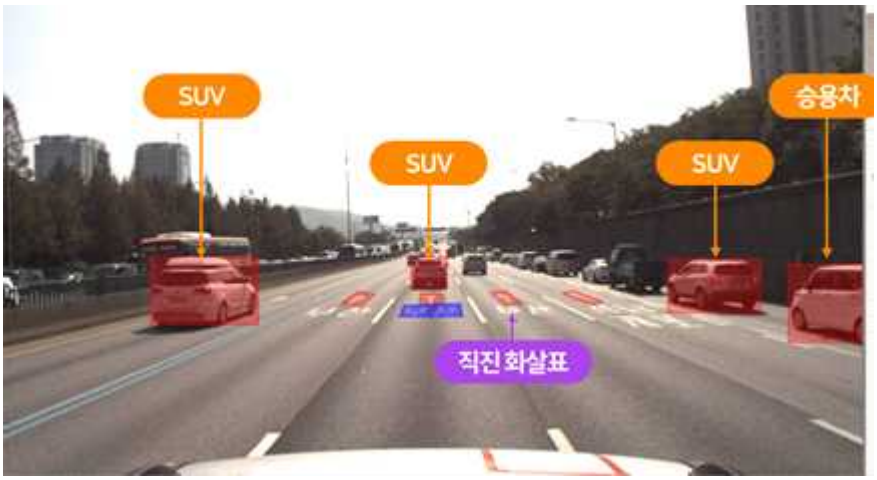
르며, 바로 이 때문에 절차상 아무 문제가 없었음에도 성별에 따른 이동통신 이용 현황 정보에는 오류가 발생한다. 앞서 언급한 '데이터 상의 오류 아닌 오류'란 이러한 경우를 말한다. 유의미한 분석 결과를 만들려면 '본인이 등록한 이동통신 회선을 직접 사용하는 사람만을 걸러내어' 남녀 간의 통화기록을 비교해야 할 것이다.

결국 데이터를 본래 수집목적에서 벗어나 다른 용도로 활용하는 경우, 늘 괴리가 발생하기 마련이다. 데이터를 유의미하게 활용하려면 원래 데이터를 활용목적에 맞게 잘 바꾸어야(가공 및 정제) 한다. 이를 다르게 표현하면 '쓸 만한 데이터가 부족하다'는 말은 '활용목적에 맞게 데이터를 가공할 기술과 능력이 부족하다'는 뜻이다. 자연히 데이터 가공능력을 키우면, 쓸 만한 데이터가 늘어날 수밖에 없다.

2.2 인공지능 분야에서 데이터 가공 및 정제

앞절에서 대부분의 빅데이터는 데이터 분석 및 활용을 위해 수집되지 않았기에 이를 잘 가공하고 정제해야 활용의 폭을 넓힐 수 있음을 알았다. 그렇다면, 인공지능 기술 개발 분야에서 데이터 가공 및 정제는 어떤 역할을 수행할까?

데이터 분석이라는 차원에서는 빅데이터 분야나 인공지능 분야의 차이는 크지 않다. 다만, 정형 데이터를 분석하는 빅데이터 분야에서는 분석을 위한 데이터 세트 구축 시 인간의 수작업(手作業)이 굳이 필요하지 않다. 그러나 비정형 데이터를 다루는 인공지능 분야에서는 인간 노동력의 투자가 선행돼야 한다. 예를 들어, 도로 및 자동차 영상 정보를 이용하여 자율주행 인공지능 알고리즘을 개발한다고 하자. 이 경우, 개발초기에는 인공지능이 무엇이 사람이고 무엇이 자동차인지 알 수 없다. 인공지능이 도로의 다양한 개체를 하나하나 인식할 수 있게 하려면 영상에서 자동차, 사람 등의 부분만 따로 추출해 학습시켜야 한다. [그림 2]는 도로 영상에서 자동차의 차종, 도로상의 표시 등을 별도로 추출하고 라벨링한 예시다.



※출처: 자율주행 차량을 위한 인공지능 학습데이터 구축, 올포랜드 컨소시엄, 2020

[그림 2] 자율주행 인공지능 알고리즘 개발을 위한 도로 정보 추출

문제는 이러한 인간의 노력이 가급적 최소화 되어야 한다는 것이다. 문제는 시간과 비용이다. 많은 양의 데이터를 미리 잘 가공해서 인공지능을 학습시키는 데 사용한다면 분명 성능 향상에는 긍정적인 영향을 준다. 그러나 이 경우 데이터의 양에 비례해 사람의 노동력과 시간이 투

입되는 만큼 적지 않은 비용이 들어 현실적으로 적용하기 어려워진다. 따라서 성능과 비용을 고려하여 타협점을 찾아야 하는데, 그중 하나가 바로 '데이터 가공의 자동화'다. 앞서 예로 든 자율주행의 경우 인공지능 개발 초기에는 자동차 차종식별을 위해 영상에서 차량 추출 작업을 인간이 일일이 수작업으로 진행했다. 그러나 어느 정도 라벨링된 영상 자료가 충분히 축적되면 가공 및 정제 규칙을 알고리즘화하여 새로 수집된 영상을 자동으로 어노테이션하고 라벨링할 수 있다. 이러한 자동화가 가능한 시스템을 얼마나 빨리 구축하는가가 성능과 비용이라는 두 마리 토끼를 한꺼번에 잡는 관건이다.

2.3 데이터 가공 및 정제 결과 활용

- 데이터 서비스 사업을 잘 할 수 있는 것은 누구인가?

현대의 정보통신 환경에서는 데이터와 아이디어만 있으면 어렵지 않게 사업모델을 만들 수 있다. 고등학교 2학년 때 개발한 '서울버스'라는 스마트폰 앱(응용프로그램)이 출시 한 달 만에 4만건 넘게 다운로드되어 화제가 된 유주완 씨(연세대)의 경우가 좋은 사례다. 유씨는 서울시에서 공개한 공공정보인 '버스 승하차 데이터'를 사용자의 눈높이에 맞춰 가공하고 정제하여 서비스함으로써 폭발적인 인기를 누렸다. 간단하지만 편리하면서도 정보 이용자의 니즈에 정확히 부합한 앱이었다[2].

이처럼 데이터를 이용한 사업모델에는 거대한 사업적 인프라나 인력이 필요하지 않다. 오히려 우리의 삶을 편안하게 할 다양한 아이디어를 발빠르게 실행하는 것과, 급변하는 라이프스타일에 맞춰 시의성을 잘 갖추는 것이 사업의 성패를 좌우한다. 생활 구석구석에서 소소하게 편리함을 느낄 수 있는 사업 아이템 상당수가 아이디어를 신속하게 사업화하는 실행력을 갖춘 가벼운 조직에서 탄생한 것도 우연이 아니다.

데이터를 활용한 서비스는 오히려 대기업이 수행하기에 제약이 많다. 조직이 크고 무거워서 의사결정과 실행이 느린 편인 데다, 소소한 편의성을 제공하는 아이디어를 다양하게 발굴하기도 어렵다. 생활 곳곳에서 참신한 편의를 제공하는 사업모델은 시장이 크지 않은 경우가 많은데, 사업 단위가 큰 대기업으로서는 매력적인 시장이 아니기 때문이다. 따라서 데이터를 활용한 사업은 창의력과 아이디어가 있는 새로운 스타트업들이 많이 생겨날 수밖에 없는 분야다.

최근에 통신이나 카드회사처럼 우리 삶과 밀접한 관련이 있는 여러 기업에서 각자 보유한 데이터를 상업적으로 활용하는 방법을 모색하고 있다. 그러나 축적된 정보를 구체적으로 어디에 어떤 식으로 사용할지에 대해서는 아이디어 빈곤에 시달리곤 한다. 이러한 상황에서는 데이터가 많은 대기업과 아이디어와 실행력이 풍부한 스타트업의 조화가 좋은 돌파구가 될 수 있다.

- 중소기업/소상공인의 데이터 활용이 어려운 이유는?

데이터를 원하는 형태로 적절히 가공하려면 데이터를 자유자재로 바꿀 수 있는 데이터 기술자(SQL 기술자), 서로 다른 형태의 데이터를 같은 데이터로 변환할 수 있는 기술자(데이터 표준화 기술자) 등이 필요하다. 여기에 더해 데이터가 원천적으로 어떻게 수집됐는지 이해할 수 있는 해당 분야별 전문가가 필요하다. 예컨대 금융 데이터를 가공해야 하는데 금융 분야의 용어나 데이터의 특성, 수집 목적 등을 모른다면, 데이터 가공이 제대로 될 수 없다.

앞서 언급한 여러 데이터 가공 전문가를 모두 갖추고 데이터 가공을 제대로 할 수 있는 기업이나 기관은 현재 국내에 많지 않다. 인력풀을 풍부하게 갖추기 어려운 중소기업에게는 특히나 어려운 일이다.

다만 데이터 가공이 필요하다고 해서 기업이 데이터 가공에 필요한 역량을 온전하게 갖출 필요는 없다는 점을 유념할 필요가 있다. 쇠고기스테이크를 먹기 위해 집집마다 도축장을 만들 필요가 없는 것처럼, 데이터 가공을 전문으로 다루는 기업이나 기관을 이용하는 것이 합리적이다. 데이터 가공 전문인력을 양성하려면 최소한 3~5년 이상의 기술교육과 실전 경험이 필요하니 산업 생태계 전체를 보아도 데이터 가공을 전문화하는 것이 효율적이라 판단된다.

3. 맺음말

우리나라는 반도체, 기계, 자동차 등 첨단산업에서 세계적 경쟁력을 확보하고 있다. 그러나 4차 산업혁명의 근간인 데이터 산업 생태계는 아직 가내수공업적 낙후성에서 벗어나지 못하고 있다. 가내수공업적 데이터 경제란 하나의 조직 혹은 기관에서 데이터 수집, 가공 및 활용, 서비스가 동시에 수행되어 데이터 관련 기술이 세분화, 전문화, 고도화되지 못하고 있다는 의미이다. 특히 인공지능에 필요한 데이터 가공은 다른 나라에 비해 전문인력과 기술 수준이 많이 뒤처진 편이다. 이를 보완하려면 새로운 데이터 가공기업을 육성함으로써 많은 중소기업이 양질의 데이터 재료를 활용해 여러 서비스와 상품을 새롭게 개발하는 데이터 산업 생태계를 조성해야 한다. 이러한 노력의 끝에는 '세계시장의 4차 산업혁명을 주도하는 대한민국'이라는 결실이 있을 것이다.

[참고문헌]

[1] 과학기술정보통신부, 대한민국 정책브리핑 '인공지능(AI)국가전략', 2019

[2] 한겨레, http://www.hani.co.kr/arti/PRINT/608840.html, 2013

※ 출처: TTA 저널 제192호

(코로나 이슈로 각 표준화기구의 표준화회의가 연기・취소됨에 따라 TTA 저널로 대체합니다)