언론진흥재단, 뉴스트러스트 알고리즘 개발과 공개 의미 : 정보 편향 막아줄 알고리즘, 소스 공개로 투명성도 확보

신문과방송 · 2018. 6. 13. 16:00

아래 사진은 2018년 5월 11일 오전 11시 30분 구글 뉴스와 네이버 뉴스의 정치 섹션 상단 주요 기사 두 개씩을 갈무리한 것이다. 같은 기사는 아니지만 북미 정상회담 관련 내용이 가장 중요한 기사로 배치돼 있는 것을 볼 수 있다.

그런데 두 번째 중요한 기사로 배치된 내용은 각각 다르다. 구글 뉴스에서는 홍대 누드모델 몰카 사건 기사가, 네이버 뉴스에서는 유승민 바른미래당 공동대표의 발언 기사가 배치돼 있다. 뉴스를 보는 관점이 서로 다르기 때문에 당연한 결과일 수도 있다. 신문의 1면이 모두 같지 않은 이유이기도 하다. 그런데 둘의 공통점이 있다. 왼쪽 사진의 빨간 박스와 오른쪽 사진의 빨간 박스 속 내용은 구글과 네이버가 해당 섹션의 기사 배치와 관련해 설명한 글이다. 구글은 “이 페이지의 기사 선별 및 게재 위치는 컴퓨터 프로그램에 의해 자동으로 결정됩니다”라고, 네이버는 “헤드라인뉴스와 각 타이틀은 기사 내용을 기반으로 자동 추출됩니다”라고 설명하고 있다. 구글과 네이버의 공통점은 기사를 자동으로 배열하는 것이다. 이 자동 배열의 원칙이 알고리즘이다.

구글 뉴스의 정치 주요 뉴스 출처-구글 캡처

네이버 뉴스의 정치 헤드라인 뉴스 출처-네이버 캡처

네이버와 구글 뉴스 배치가 다른 이유

현재 우리가 일상적으로 사용하고 있는 컴퓨터는 하드웨어와 소프트웨어로 구성된다. 컴퓨터의 하드웨어는 사칙연산과 논리연산만 수행하는 단순한 기계에 불과해 소프트웨어가 없으면 고철 덩어리나 다름없다. 하드웨어가 연산을 통해 수행할 일을 순서대로 알려주는 명령어의 집합이 알고리즘이며, 이 알고리즘을 기계가 이해할 수 있는 언어로 표현한 것이 프로그램이다. 소프트웨어는 이러한 프로그램과 프로그램이 구동하기 위해 필요한 데이터와 관련한 문서들로 구성된다. 앞에 예로 든 사진 속 기사 배치는 알고리즘에 따라 컴퓨터가 자동으로 수행한 것이다.

컴퓨터가 수행할 일을 지시하는 명령어의 집합으로서 알고리즘은 정확하고 분명해야 한다. 따라서 이 알고리즘의 기본 동작은 매우 단순하다. 논리곱(AND), 논리합(OR), 부정(NOT)이다. 논리곱, 논리합, 부정 세 가지 기본 동작이면 아무리 복잡한 알고리즘도 표현이 가능하다. 횟수를 충분히 늘리면 된다. 그런데 기계적 방식에 따른 구글과 네이버의 배열 결과는 왜 서로 다를까?

알고리즘은 데이터를 처리하는 규칙이다. 어떠한 데이터가 주어졌을 때 어떻게 처리하라는 규칙의 집합으로 그 데이터를 처리하는 전략이라고 할 수 있다. 이 전략은 수립하는 사람 혹은 집단에 따라 다르다. 각자가 갖고 있는 지식과 논리에 따라 전략이 도출되기 때문이다. 구글과 네이버의 기사 배열 결과가 서로 다른 첫 번째 이유다. 구글과 네이버는 어떠한 기사를 더 비중 있게 다룰지 등에 대해 서로 다른 전략을 적용하고 있다. 두 번째는 기사 배열 전략의 기본이 되는 데이터, 즉 수집한 뉴스가 다르기 때문이다.

기사는 사람이 쓴다. 수많은 언론사의 기자들이 쓰는 기사는 그 수만큼 형식도 다양하다. 알고리즘은 정해진 유형의 데이터를 처리할 때는 강력하지만, 뉴스의 형식은 사람마다 다른 만큼 굉장히 다양하다. 고려해야 할 요인들이 너무 많고 각 요인들은 복잡하게 뒤얽혀 있다. 같은 알고리즘을 적용하더라도 처리하는 데이터에 따라 다른 결과를 가져올 수 있다. 수많은 형식의 다양한 데이터를 처리하기 위해 알고리즘도 그만큼 복잡해질 수밖에 없다. 이러다 보니 현재 알고리즘은 한 사람이 이해할 수준을 넘어섰다. 알고리즘이 복잡해지면서 문제가 생겨도 사람이 발견하지 못하는 경우도 자주 발생한다.

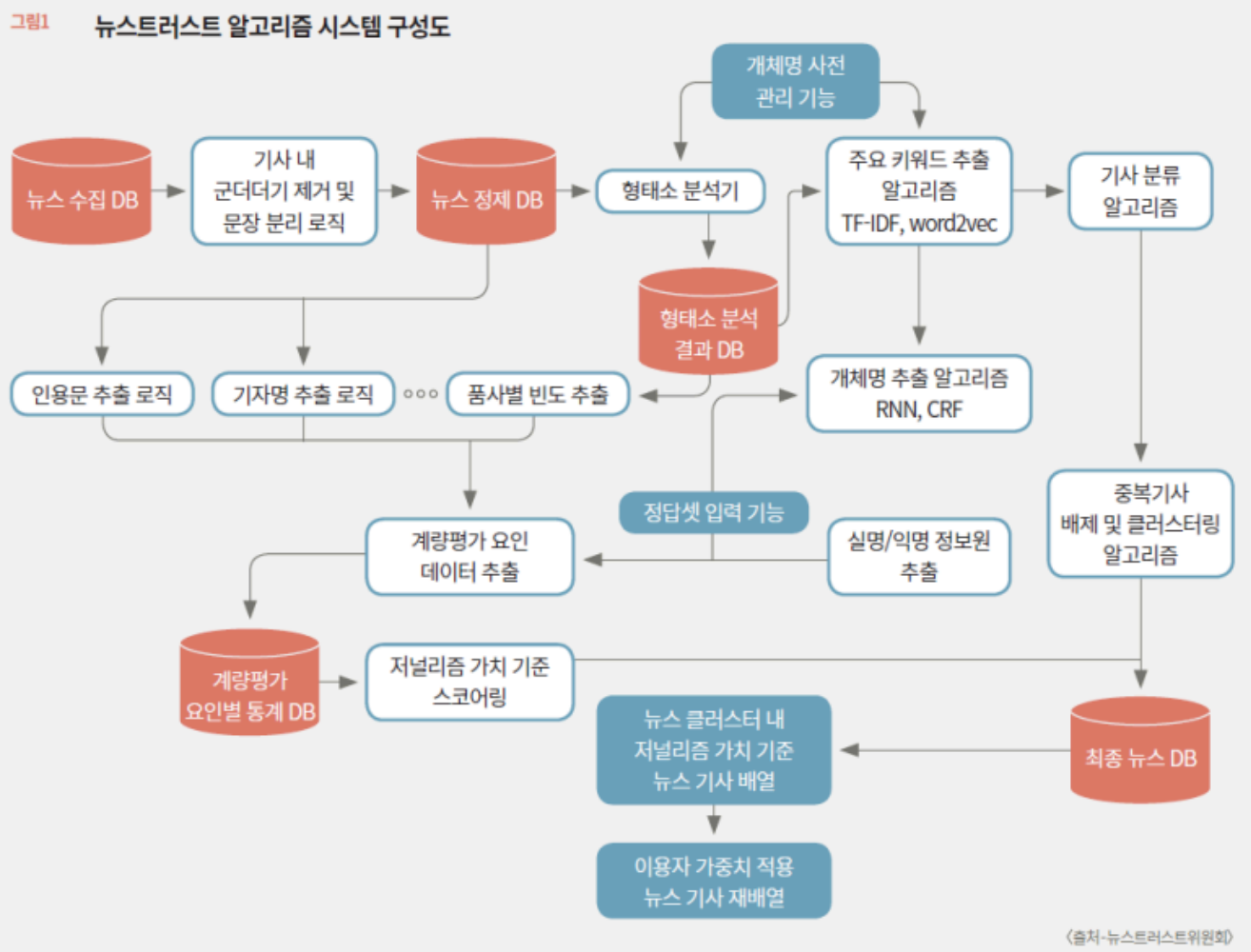
저널리즘 가치에 기반한 알고리즘

지난 5월 구글과 네이버는 100% 알고리즘에 의한 맞춤형 뉴스를 강화하고 사람의 편집을 없애겠다는 방침을 발표했다. 여러 가지 이유가 있겠지만, 뉴스에 대한 선호가 각기 다른 이용자들이 관심 있어 하는 뉴스를 위주로 배열하겠다는 전략이다. 각 개인들이 보고 싶어 하는 뉴스 중심으로 전달해 더 많은 뉴스 소비를 이끌어내고 필요한 뉴스를 효율적으로 전달하겠다는 것이다. 음악, 영화, 책 등의 경우 소비가 늘어나는 것이 큰 문제가 되지 않는다. 맞춤형 콘텐츠 추천은 이 경우 문제가 없다. 하지만 뉴스라는 점에서 문제가 복잡해진다. 건강한 민주주의 공동체 유지와 발전을 위해서는 우리 사회에 다양한 의견이 필요하다. 싫은 목소리도 들어야 한다. 보통 싫은 목소리는 뉴스를 통해 전달된다. 맞춤형으로 누군가가 보고 싶어 하는 뉴스만 전달하다 보면 듣기 싫은 목소리를 듣지 못하게 될 가능성이 크다. 최소한 나와 다른 의견이 있다는 사실이라도 알아야 하는데, 맞춤형 뉴스 추천 알고리즘은 그 목적상 개인이 그동안 관심 없어한 뉴스는 전달하지 않을 가능성이 높다. 뉴스를 많이 읽는 것만큼이나 다양한 의견을 듣는 것도 중요하다.

맞춤형 뉴스 추천 알고리즘이 의견 다양성 측면에서 가질 수 있는 문제, 어떤 과정을 거쳐 결과를 제시하는지 알 수 없는 알고리즘의 복잡성 문제 등이 있지만 맞춤형 뉴스 추천 알고리즘은 계속해서 개발과 적용이 이어질 것이다. 이용자 입장에서 뉴스 소비의 편의성만 놓고 봤을 때는 효율적이기 때문이다. 이것이 뉴스트러스트위원회 (위원장 김춘식)의 출범 배경이다. 효율적이지만 문제도 갖고 있는 뉴스 추천 알고리즘을 보완할 수 있는 대안적 알고리즘을 개발해보자는 것이다. 뉴스트러스트위원회는 “페이지뷰

나 트래픽을 위한 알고리즘이 아니라, 식견 있는 공중을 위해 최적화된 알고리즘을 설계하고 구현할 수 있을까"(Lotan, 2014, p.118)[1]라는 질문에서 출발했다. 뉴스 이용자의 선호도가 아니라 저널리즘 가치에 기반 둔 뉴스 추천 알고리즘을 개발해 맞춤형과 달리 뉴스 배열에서 의견의 다양성을 담아내고자 한 것이다. 또한 자연어 처리, 형태소 분석, 클러스터링 등 알고리즘 원천 기술을 오픈소스 방식으로 모두 공개하고 알고리즘의 기본 원칙 등도 가능한 상세하게 공개해 알고리즘의 투명성도 제고해보고자 했다.

그림1 뉴스트러스트 알고리즘 시스템 구성도



지난 4월 12일 뉴스트러스트위원회는 현재까지 개발한 뉴스트러스트 알고리즘의 중간 결과물[2]을 발표했다. 지난 2016년 5월 4일 위원회 출범 이후 주요 결정 사항, 뉴스의 저널리즘 가치, 알고리즘 기술(형태소 분석, 클러스터링, 자동분류, 중복 필터링, 개체명 인식 등), 뉴스 측정 계량 요인별 점수 부여 공식, 계량 요인 가중치, 뉴스 랭킹 산출 공식 등 대부분의 내용을 공개했다. 특히 일종의 설계도라고 할 수 있는 소스 코드도 깃허브[3]에 공개해 원하는 사람들은 누구나 뉴스트러스트 알고리즘을 이용하고 검증해 볼 수 있도록 제공했다. [그림1]과 같이 방대한 내용을 모두 소개하는 것은 불가능하기 때문에 그중 핵심적인 일부로서 뉴스트러스트 알고리즘의 뉴스 계량 평가 요인과 그 적용 공식에 대해서만 소개한다.

뉴스 계량 평가 요인과 적용 기준

뉴스트러스트 위원회는 기자 이름, 기사의 길이, 인용문 수, 제목 길이, 제목의 물음표, 느낌표 수, 수치 인용 수, 이미지 수, 평균 문장 길이, 제목의 부사 수, 문장당 평균 부사 수, 기사 본문 중 인용문 비중 등 11개 요인만 우선 뉴스트러스트 알고리즘에 적용했다. 11개 요인만을 적용한 이유는 현재 기술 개발 진행상 정확도가 90% 이상인 것으로 한정했기 때문이다. 뉴스트러스트위원회는 11개 요인 외에 추가 요인을 개발 중이며, 추가 개발 사항도 향후 상세하게 공개할 예정이다.

한편 11개 요인 중 일부는 플랫폼 기업들의 뉴스 배열 알고리즘에도 이미 적용되고 있다. 뉴스트러스트 알고리즘 공개가 의미를 지니는 것은 이러한 11개 요인의 구체적인 적용 방식을 밝혔기 때문이다. 예를 들어 기사의 길이 같은 경우 구글 등 플랫폼 기업들도 기사 배열에 있어 중요한 요인으로 적용하고 있다. 하지만 “기사의 길이가 길면 더 높게 평가한다” 등 기본 원칙만 제시할 뿐 얼마나 길어야 하는지, 길이에 따른 랭킹은 어떻게 부여하는지와 같은 구체적인 기준은 제시하지 않고 있다.

> “
> **뉴스트러스트 알고리즘은**
> **기준 및 가중치는 물론 소스 코드까지**
> **모두 공개했다.**
> **이는 뉴스트러스트위원회가 정답이 아니며,**
> **향후 관심 있는 모든 사람들과 함께**
> **알고리즘을 논의해 더 나은 대안을**
> **만들어가겠다는**
> **의지의 표현이라고 할 수 있다.**
> ”

뉴스트러스트위원회는 알고리즘 중간 공개를 통해 11개 요인의 적용 기준을 구체적으로 공개했다. 각각 요인을 대상으로 ①저널리즘적 의미 ② 조작적 정의 ③추출 방법 ④평가 적용 방식 등으로 구분해 제시했다. 예를 들어 기사의 길이가 길수록 사건과 관련한 완전한 내용(whole story)을 다루고 있을 가능

성이 높으며, 정보량이 많고 심층성이 있으며, 다루는 정보의 범위가 넓어 좋은 기사로 판단할 수 있다고 저널리즘적 의미를 설명한 뒤 제목, 소제목, 사진 설명 등을 제외한 기사 본문의 길이라고 조작적으로 정의했다. 또한 제목, 관련 기사, 광고 등 기사 본문 외 데이터를 정제하고 'utf-8' 유니코드 기준으로 기사 본문을 원고지 글자 수로 길이를 계산했다고 추출 방법을 제시했다.

가장 중요한 것은 이렇게 추출한 기사의 길이를 계량화하는 기준을 제시한 점이다. 뉴스트러스트 위원회는 길이가 긴 것에 대한 보상이 있어야 한다는 관점에서 길이에 따라 0~1점 사이의 가점을 부여했다. 좀 더 구체적으로는 기사의 분류, 매체 유형(신문 · 방송) 등에 따라 기사의 길이 평균을 추출한 후 평균 이상인 기사에 대해 표준편차에 따라 단계별 가점을 부여했다. 그 결과 기사의 길이는 0, 0.165, 0.33, 0.495, 0.66, 0.835, 1 등 7단계의 값을 갖게 되었고, 이를 코드로 표현하면 다음과 같다.

```
if (content_length < mean) then 0,
        else (content_length   < mean + 0.5SD) then 0.165,
        else   (content_length < mean + SD) then 0.33,
        else   (content_length < mean + 1.5SD) then 0.495,
        else   (content_length < mean + 2SD) then 0.66
        else   (content_length < mean + 2.5SD) then 0.835
        else (content_length   > mean + 2.5SD) then 1
```

기자 이름의 경우 기자가 직접 쓴 기사는 믿을 만하다(trustworthy)고 판단할 수 있으며, 기명이 아닌 경우는 기사 작성 과정에서 어뷰징 등 비정상적 요인의 개입이 있었다고 판단할 수 있다는 점을 감안해 뉴스트러스트위원회는 기사 본문 내 명기돼 있는 기자 이름을 최대한 추출했다. 추출 방법은 기자 이름 DB 필드에 이름이 있을 경우 이를 통해 추출했고, DB 필드에는 없으나 본문에 이름이 있는 경우 언론사별 특성을 일일이 파악해 기계적으로 추출했다. 추출된 기자 이름 중 인터넷뉴스팀과 같은 실명이 아닌 경우 비실명 기자 이름으로 별도 처리했다. 이후 기사 분류 및 매체 유형과 상관없이 실제 기자 이름과 이메일이 병기돼 있을 경우에는 1점의 가점을 부여하고, 기자 이름만 있으면 0.8점, 비실명 기자 이름일 경우에는 0점, 기자 이름 없이 이메일만 있을 경우에는 0점을 부여했다. 기자 이름과 이메일이 아예 없을 경우에는 -1점의 감점을 부여했다. 기사의 길이, 기자 이름 외 9개 요인에 대해서도 이와 같은 구체적인 기준을 제시했으며, 자세한 내용은 뉴스트러스트위원회가 발표한 설명 자료4를 참고하면 된다. 한편 이러한 기준에 정답은 없다고 할 수 있기 때문에 이것이 확정된 것은 아니며, 공개 내용에 대한 피드백 등을 받아 향후 지속적으로 수정해나갈 예정이다.

저널리즘 가치별 가중치 부여

기본 값을 추출한 이후에는 저널리즘 가치에 따른 가중치를 부여했다. 뉴스트러스트위원회는 균형성, 다양성, 독이성, 독창성, 사실성, 심층성, 유용성, 중요성, 투명성, 선정성 등 10개의 저널리즘 가치를 현재 디지털 뉴스 영역에서 중요한 가치라고 판단했다. 이후 11개의 계량 요인들이 각 저널리즘 가치에 따라 갖는 가중치를 이론적으로 도출했다. 예를 들어, 기사의 길이는 심층성과 관련해 4~5, 사실성과 관련해서는 3.5~4.5의 가중치를 갖는다고 위원회 논의를 통해 결정했다. 가중치를 고정된 값이 아닌 구간으로 설정한 이유는 기계 학습을 위한 여지를 두기 위해서다. 뉴스트러스트위원회는 각 기사에 대해 위원들이 개별적으로 평가하도록 했다. 위원들은 저널리즘 가치에 따라 각 기사들을 평가해 점수를 부여했다. 그 결과는 알고리즘이 맞춰야 할 목표로 설정했다. 뉴스트러스트위원회는 위원들의 평가 내용을 정답 셋으로 해서 위원들이 이론적으로 도출한 저널리즘 가치에 따른 계량 요인의 가중치 구간의 정확한 값을 찾도록 기계 학습을 시켰다. 즉 사람이 저널리즘 가치에 따른 적절한 구간을 제시한 후 기계가 그 안에서 정확한 값을 찾도록 한 것이다. 이는 기계 학습에 의해 특정 요인에 치중된 가중치 값을 갖게 될 수도 있는 부분을 제어하기 위한 목적이었다. 그 결과 도출된 기사의 길이와 기자 이름의 저널리즘 가치별 가중치는 [표]와 같다.

표 **저널리즘 가치에 따른 11개 계량 요인들의 가중치**

구분	위원회의 이론적 가중치		기계 학습에 따른 가중치	
	기자 이름	기사의 길이	기자 이름	기사의 길이
균형성	2~3	3~4	2.996	3.002
다양성	0~1	4~5	0.998	4.994
독이성	0~1	0~1	0.001	0.003
독창성	3.5~4.5	3.5~4.5	4.494	4.492
사실성	3.5~4.5	3.5~4.5	4.493	3.503
심층성	3.5~4.5	4~5	4.496	4.995
유용성	2.5~3.5	2.5~3.5	3.494	3.498
중요성	1.5~2.5	3.5~4.5	2.495	3.503
투명성	3.5~4.5	3~4	4.498	3.003
선정성	3.5~4.5	0	4.491	0

•11개 계량 요인 중 이 표에서는 '기자 이름'과 '기사의 길이'에 대한 가중치만 제시했음. 〈출처-뉴스트러스트위원회〉

이러한 가중치까지 산출한 결과, 개별 기사들은 다음과 같이 세 가지 값을 갖게 된다. ①계량 요인에 따른 기본 값으로, 각 계량 요인에 따라 산출된 값의 합이다. 예를 들어, '기자 이름(0.8)+기사의 길이(0.66)+제목의 부사 수(-1)+인용문 비중(-0.5) =-0.04'와 같은 방식으로 각 개별 기사가 갖는 고유의 값이다. ②각 요인의 저널리즘 가치별 값의 합으로, 각 계량 요인이 해당하는 저널리즘 가치별로 값을 적용한 후 값이다. 예를 들어, '독이성 =기사 길이(1)+이미지 수(0.66)+평균 문장 길이(0)+제목의 부사 수(-0.5)=1.16' 등과 같은 계산 방식에 따른 결과 값이다. ③저널리즘별 가중치 적용 후 합산 값으로, 각 계량 요인에 따라 산출된 값에 저널리즘 가치별 가중치를 적용한 값의 합이다. 예를 들어, '기사의 길이(0.66)X {균형성(1)+다양성(1.17)+독이성(1)+독창성(1) +사실성(1)+심층성(1.67)+유용성(1)+중요성

(1.25)+투명성(1)}=6.6594', '제목의 부사수(0.5)× {독이성(1)+선정성(1.3)}=-1.15' 등과 같은 식으로 저널리즘 가치별 가중치가 적용된 값들의 합이다. 이렇듯 뉴스트러스트 알고리즘에 따르면 각 개별 기사는 세 가지의 값을 갖게 되며, 이 값은 각 기사마다 고정된 값이다.

뉴스 배열과 관련한 알고리즘 중 구체적 기준과 가중치까지 상세하게 공개하는 경우는 뉴스트러스트 알고리즘이 처음이라고 할 수 있다. 기업이 만든 알고리즘의 경우 일종의 영업 노하우로 저작권 보호 대상인 동시에 어뷰징 등 악의적 이용 우려도 있어 공개가 어렵기 때문이다. 하지만 뉴스트러스트위원회는 공익적 관점에서 대안 모색을 목적으로 했기 때문에 이렇듯 상세한 공개가 가능했다. 또한 이러한 기준 및 가중치 공개와 함께 소스 코드까지 모두 공개해 원하는 개인, 기업 등 누구나 시험 및 적용할 수 있다. 이는 뉴스트러스트위원회가 정답이 아니며, 향후 관심 있는 모든 사람들과 함께 알고리즘을 논의해 더 나은 대안을 만들어가겠다는 의지의 표현이라고 할 수 있다. 이러한 관점에서 뉴스트러스트위원회는 이용자들이 저널리즘 가치에 따른 가중치를 별도로 설정해서 기사 배열을 달리할 수 있는 기능도 추가로 개발했다. [그림2]와 같이 이용자들은 균형성, 심층성, 사실성 등 위원회 논의를 통해 정해진 10개의 저널리즘 가치에 대해 본인이 원하는 정도를 0~4단계로 설정해 배열 결과를 다르게 할 수 있다.

계량 요인 늘려갈 예정

이번에 공개한 뉴스트러스트 알고리즘은 완성품이 아닌 목표를 향해 가는 가운데 발표하는 중간 결과

다. 뉴스 배열 알고리즘의 100% 완성품은 사실 없다고 할 수 있으며, 뉴스트러스트위원회의 알고리즘은 저널리즘 가치 기반의 뉴스 배열 알고리즘 개발이라는 목표를 향해 가는 중간 과정이라 할 수 있다. 현 단계 11개 계량 평가 요인은 뉴스트러스트위원회가 목표로 한 평가 요인 중 현재 추출 정확도가 90% 이상인 것만을 발표하는 것으로 3차 연도 개발을 통해 계량 요인(실명 정보원·익명 정보원, 무주체 술어 등) 수를 늘려나갈 예정이다. 또한 소스 코드 공개를 통해 언론사, 일반 이용자 등의 다양한 의견을 수렴해 향후 개발에 적극적으로 반영할 예정이며, 2018년 말에는 현재의 설명 자료보다 상세한 백서를 발간해 뉴스트러스트위원회가 개발 과정에서 겪은 고민, 시행착오, 어려움 등을 공유할 예정이다.

한편 뉴스트러스트위원회는 모든 것을 공개한다는 목표를 세웠지만, 저작권 문제로 인해 뉴스트러스트 알고리즘 개발에 활용한 뉴스 기사 데이터의 경우 일반 공개를 못하고 빅카인즈 제휴 언론사 45곳에만 제한적으로 공개할 수밖에 없었다. 빅카인즈 제휴 언론사 외 언론사와 일반 국민은 자체적으로 뉴스 기사를 확보한 후 공개된 소스 코드를 활용해 시스템을 구축해야 뉴스트러스트 알고리즘을 활용할 수 있다. 기계 학습용 사전, 기계 학습용 위원회 기사 평가 내역, 기계 학습용 정답셋 등의 경우도 저작권 문제로 인해 이번 공개 대상에서 제외됐다. 향후 뉴스트러스트 알고리즘 개발을 위해 사용한 사전, 기계 학습용 데이터 등을 모두 무료로 공개해 사회 자원화하고, 개선할 수 있는 방법론을 제시하기 위해 문제 해결에 적극 나설 예정이다.

-

1 Lotan, G. (2014), Networked audiences: Attention and data-informed. In K. McBride, & T. Rosenstiel, (Eds.), The New Ethics of Journalism: Principles for the 21st Century. Thousand Oaks, CA: CQ Press, Sage. (pp. 105-122)

2 http://www.kpf.or.kr/site/kpf/ex/board/View.do?cbIdx=246&bcIdx=20452

3 https://github.com/KPF-NEWSTRUST

4 http://www.kpf.or.kr/site/kpf/03/10309000000002016110402.jsp

글 / 오세욱 (한국언론진흥재단 선임연구위원· 뉴스트러스트위원회 위원)

- 본 기사는 <신문과방송> 2018년 6월호(통권 570호) 산업·정책 섹션에 수록되어 있습니다. -