저널리즘 가치에 기초한 알고리즘을 이용한 뉴스 시각화

A news visualization based on an algorithm by journalistic values

박대민, Daemin Park*, 김기남, Gi-Nam Kim**, 강남용, Nam-Yong Kang***,
서봉원, Bongwon Suh****, 하효지, Hyo-Ji Ha*****, 온병원, Byung-Won On******

요약 현재 온라인 뉴스 서비스는 선정적인 연성뉴스 중심으로 제공된다. 이에 따라 저널리즘 가치를 구현한 뉴스 서비스의 필요성이 대두되고 있다. 정보원과 공동 인용 여부에 따라 기사를 클러스터링하고 가중치를 부여해 사실성, 다양성, 심층성, 비판성 등 주요 저널리즘 가치를 구현한 알고리즘은 뉴스정보원연결망분석(news source network analysis)으로 제안된 바 있다. 본 연구는 이를 사용자 친화적으로 시각화한 서비스인 뉴스소스를 제안한다. 뉴스소스는 시간과 정보원에 따라 뉴스를 막대그래프 형식으로 어떤 토픽에 대해 분야별, 소속별로 얼마만큼의 중요도에 따라 논의되는지를 대조적으로 보여준다. 본 연구에서는 뉴스 아카이브인 카인즈의 기사를 활용해 뉴스소스의 베타 버전을 구현했다.
(http : //147.47.125.161/NSNA/ 에서 베타서비스 중이며, 구글 크롬에 최적화 되어있음)

Abstract There was widespread criticism of the online news services due to their bias toward sensational and soft news. Thus, news services based on journalist values are socially requested. News source network analysis(NSNA), an algorithm to cluster and weight news sources, quotes, and articles, is suggested as a method to emphasize on journalist values like facts, variety, depth, and criticism in the previous study. This study suggests 'News Sources' as a visualization tool of NSNA. 'News Sources' shows news as bar graphs, weighted by facts and criticism, and arranged by organizations and subjects. This study designed a beta version using KINDS, a news archive of Korean Press Foundation.

핵심어 : News Source, news source network analysis (NSNA), journalistic values, facts, variety, depth, criticism, KINDS

이 연구는 2014년 HCI KOREA 2014 ST4 가시화 세션에서 발표된 논문 '저널리즘 가치에 기초한 알고리즘을 이용한 뉴스의 시각화'를 수정한 것임. 이 연구는 2013년도 미래창조부와 한국정보화진흥원 빅데이터 활용 스마트서비스 시범사업의 지원을 통하여 연구되었음.

* 책임저자 : 한국언론진흥재단 연구위원; e-mail : heathe0@gmail.com
** 아주대학교 라이프미디어협동과정 석사과정; e-mail : gnkim@ajou.ac.kr
*** 아주대학교 정보컴퓨터공학과 학사과정; e-mail : fnantia@ajou.ac.kr
**** 서울대학교 융합기술과학기술대학원 융합과학부 부교수; bongwon@snu.ac.kr
***** 아주대학교 라이프미디어협동과정 석사과정; e-mail : hjha0508@ajou.ac.kr
****** 교신저자 : 군산대학교 통계컴퓨터과학과 조교수; e-mail : bwon@kunsan.ac.kr

■ 접수일 : 2014년 3월 16일 / 심사일 : 2014년 4월 5일 / 게재확정일 : 2014년 8월 26일

1. 서론

현재 포털 뉴스 검색 서비스에서는 연예, 스포츠, 성(sex), 범죄, 인물, 미담 등 선정적 기사 또는 연성뉴스(soft news)가 중시되는 황색 저널리즘(yellow journalism)의 성격이 강하게 나타난다[1]. 포털 저널리즘이 인터넷 담론 공중[2]의 토론 기초가 되는 등 긍정적 기능이 있음에도 불구하고, 적지 않은 비판을 받는 것도 이 때문이다[3]. 이는 광고 수익 극대화를 위해 클릭 수를 늘리려는 언론사의 온라인 뉴스 편집 관행 때문이기도 하지만, 포털 역시 실시간 검색어 등 인기도에 기초해 뉴스 가중치를 부여하는 등 저널리즘 가치와는 무관한 알고리즘을 적용하는 탓도 있다.

저널리즘의 주요 가치로는 사실성, 다양성, 심층성, 비판성 등이 널리 인정받는다. 우선 사실성 관행으로는 대표적으로 인용, 수치, 사례를 꼽을 수 있다[4]. 이 가운데 인용은 가장 중요하다. 특히 인용 대상인 정보원은 기사 내용의 핵심 제공자로 저널리즘 연구자들이 일찍부터 주목했다[5], [6], [7].

정보원과 공동 인용 여부에 따라 기사를 클러스터링하고 가중치를 부여해 사실성, 다양성, 심층성, 비판성 등 저널리즘 가치를 구현한 알고리즘은 뉴스정보원연결망분석(news source network analysis, NSNA)으로 제안된 바 있다[8]. 뉴스정보원 연결망은 같은 기사에 두 정보원이 직접인용문으로 함께 인용됐을 경우 이 정보원들 간에 서로 의미론적인 관계가 있는 것으로 보고 간접적으로 만드는 양방향(undirected) 준 연결망(quasi network)이다. 본 연구는 NSNA를 사용자 친화적으로 시각화한 뉴스소스를 제안한다. 뉴스소스는 뉴스 보기 화면을 정보원들의 토론장처럼 시각화한다. 즉 어떤 토픽에 대한 분야별, 소속별 의견을 사실에 가중치를 두어 시간에 따라 대조할 수 있도록 막대그래프를 활용함으로써 저널리즘 가치를 반영한 뉴스의 시각화를 구현하고자 했다.

2. 관련 사례 및 기존 연구 검토

1) 뉴스시각화 관련 사례

최근에 나타나는 뉴스 어플리케이션은 단순하게 뉴스기사만을 전달하는 기능에서 벗어나, 한편으로는 모바일 환경에 최적화되고 다른 한편으로는 데이터 분석을 통해 새로운 가치를 창출하는 방향으로 바뀌고 있으며 이를 위해 시각화 기술을 활용하는 사례가 늘고 있다.

본 연구에서 참조한 예로는 Marcos Weskamp의 Newsmap이 있다[9]. Marcos의 Newsmap은 Google의 뉴스 서비스를 기반데이터로 Squarified Treemap Algorithm을 이용해 그림 1과 같이 실시간으로 뉴스의 클러스터링과 중요도를 각각 색상과 사각형의 크기를 통해 시각적으로 제공한다[10].

그림 1. Newsmap 웹 사이트

또 다른 사례로 NBC news에서 서비스하는 Spectra를 들 수 있다. Spectra는 사용자가 원하는 채널(분야)을 선택함에 따라 해당 채널의 뉴스 기사들이 화면 상단에서 일정한 궤도를 형성하며 회전한다[11].

그림 2. 인포테이먼트 요소를 적용한 뉴스 어플리케이션 Spectra

2) 뉴스정보원연결망분석

본 연구는 시각화를 위한 기사, 정보원, 인용문의 가중치부여에 NSNA를 활용한다. 그림 3은 뉴스정보원연결망의 한 사례이다. 각 node는 정보원을, edge는 기사공동인용을 의미한다. 그림 1에서 검색어는 '총선'이었으며 검색기간은 19대 국회의원 선거 후보자 등록이 시작된 2012년 3월 22일부터 총선을 진행한 2012년 4월 10일까지로 설정했다. 검색매체는 한겨레신문과 동아일보였다. 그림 3은 연결정도가 0~1인 정보원을 제외한 뒤 가장 큰 주요구성집단(main component)인 뉴스정보원 연결망을 UCINET을 통해 시각화한 것이다. 가장 중요한 정보원은 박근혜 당시 새누리당 중앙선대위원장이었으며, 한명숙 민주통합당 대표, 이상일 새누리당 대변인, 김유정 민주통합당 대변인, 이정희 통합진보당 대표, 박영선 민주통합당 의원 순이었다.

NSNA 알고리즘은 사실성과 비판성에 의해 가중치를 부여하는 알고리즘이다. NSNA에 의해 중시되는 기사는 논쟁적인 관련 기사가 많은 기사이며, 정보원의 경우 논쟁적인 기사에 많

이 실린 정보원이 중요하게 취급된다.

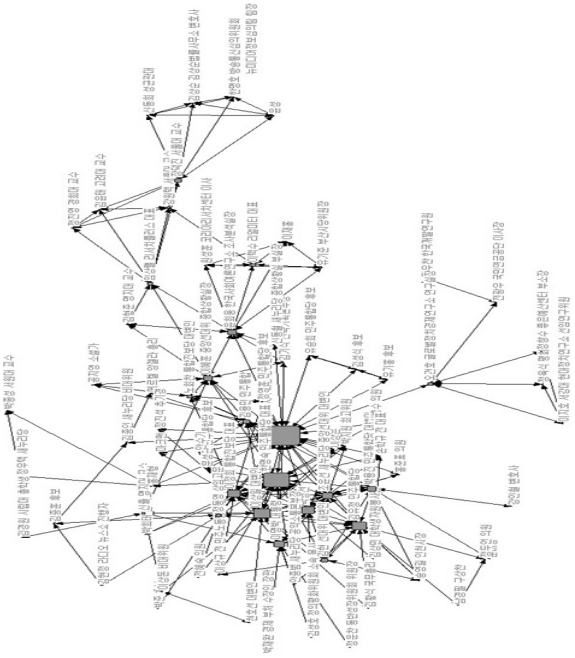

그림 3. '총선'을 검색어로 한 뉴스정보원연결망 사례

3. 뉴스소스의 구현

1) 뉴스소스 서비스 개요

뉴스소스는 NSNA 알고리즘으로 각 토픽에 대해 분야별, 소속별로 중요한 정보원을 도출하고, 정보원의 중요도에 따라 가장 중요한 기사를 막대그래프 형식의 목록으로 보여준다. 현재 한국언론진흥재단의 미디어 관련 자료 아카이브인 미디어가온(www.mediagaon.or.kr)을 통해 베타버전을 공개한 상태이다. 개발언어는 자바(JAVA)를 기본으로 했다. 자연어처리는 '꼬꼬마 형태소 분석기(kkma.snu.ac.kr)', 검색엔진은 Lucene 등을 국문 기사 분석에 맞게 보완해 사용했고, 자료 저장 및 관리와 처리에는 Linux, Hadoop, HBase, MapReduce, Hive 등의 오픈소스를 활용했다. 뉴스소스의 UI(user interface)는 웹 기반으로 PC는 물론 태블릿PC와 스마트폰 등 다양한 미디어 기기를 통해 접근이 용이하도록 HTML5로 제작됐다. 뉴스소스는 W3C CSS 레벨 3 형식의 검사를 통과했다. 일반인을 대상으로 한 뉴스소스 서비스의 기능은 크게 두 가지로 나눠진다. 첫째, '오늘의 뉴스'로 그날그날의 뉴스를 중복을 제거해 가중치에 따라 보여준다. 둘째, '검색'으로 검색어를 입력하면 검색기간별로 해당 검색어와 관련된 주요 정보원의 발언과 기사를 역시 중복없이 중요도에 따라 보여준다.

2) 사용 데이터 및 데이터 처리 방법

뉴스 데이터로는 한국언론진흥재단과의 협약을 통해 제공받은 뉴스 아카이브인 카인즈의 데이터를 활용했다. 카인즈는 포털뉴스처럼 실시간 뉴스 서비스를 제공 중이며 뉴스소스에서는 1990년 1월 1일부터 2014년 4월 10일까지 총 68개 매체 2900만 여건의 기사 데이터를 가공해 시각화했다. 향후 한국언론진흥재단과 저작권 문제 등을 협의 후 실시간 연동 서비스를 제공할 예정이다.

표 1. 뉴스소스 서비스 매체 종류

분류	매체명	계
전국종합 일간신문	문화일보, 한겨레신문, 서울신문, 국민일보, 내일신문, 경향신문, 한국일보, 동아일보, 아시아투데이, 세계일보	10
지역 일간신문	제민일보, 한라일보, 경인일보, 인천일보, 경기일보, 충북일보, 대전일보, 중도일보, 충청투데이, 중부매일, 매일신문, 경남도민일보, 경남신문, 영남일보, 경상일보, 부산일보, 국제신문, 전북일보, 전북도민일보, 새전북신문, 전남일보, 무등일보, 광주일보, 강원도민일보, 강원일보	25
경제 일간신문	서울경제, 파이낸셜뉴스, 프라임경제, 헤럴드경제, 머니투데이, 이투데이, 한국경제, 파이낸스투데이, 매일경제, 한국재경신문	10
스포츠지	스포츠서울, 스포츠칸	2
TV뉴스	SBS, MBC, KNN, KBS	4
영자신문	Korea Herald	1
인터넷 신문	대덕넷, 이데일리, 투데이코리아, 오마이뉴스, 노컷뉴스, 조세일보, 브레이크뉴스	7
지역주간지	홍성신문, 김포뉴스, 당진시대, 옥천신문	4
전문신문	국방일보, PD저널, 기자협회보, 미디어오늘	4
시사잡지	시사인	1

기사 데이터는 텍스트 형태의 비정형 데이터이다. 이를 뉴스소스 시스템에서 시각화할 수 있는 형태로 정형화하고 가중치를 부여하기 위해 자연어처리와 뉴스정보원연결망분석 등을 실시했다.

우선 기사와 정보원의 식별 과정을 설명하면, 기사를 문장 단위로 분할하고 인용문 추출한 다음 정보원의 인명・소속・직함의 추출에서는 룰과 인명사전을 활용했다. 문장 식별은 쌍따옴표나 숫자를 이용하기 때문에 거의 100% 식별된다. 인명, 소속, 직함 식별의 정확도는 일반적으로는 60~90%이지만, 사전 활용과 함께 저널리즘 영역지식으로 보강하여 룰 기반을 활용한 결과 2013년 7월 20일자 기사 기준으로 인간이 코딩한 것과 비교해 이름은 90.3%, 소속은 92.7%, 직함은 97.8%의 높은 정확도를 보였다.

다음으로 정보원 매칭(예컨대 '박원순 서울시장'과 '박 시장'의 매칭)은 SVM(support vector machine)을 활용했다[12]. 정보원 매칭 성능의 경우 무작위 선정 300개의 개체쌍에 대해 10-fold cross validation를 수행한 결과 95%의 정확도를 보였다.

중복기사를 제거하기 위한 유사 기사 및 유사 문장의 군집화(clustering)는 인명・소속・직함 및 주요 명사를 중심으로 문서 간의 코사인 유사도(cosine similarity)를 계산하여 일정 수준 이상의 유사한 기사끼리 묶어주었다. 정확도, 재현율, F-measure, 순수도 측정결과 결과 90~100%의 평가 성능을 보였다. 정보원 식별에 날짜, 성, 이름, 소속, 직함, 수치 등 6개의 정보를 핵심적으로 활용해 정확도가 개선됐다. 해당 일에 유사 기사가 없다면, 군집화 과정은 생략된다. 이는 그 주제가 그만큼 덜 논쟁적이라는 것을 의미한다. 또 이런 기사의 사실정보량은 대표기사와 유사기사의 군집보다 적을 수밖에 없다. 따라서 가중치는 대체로 낮게 평가되며 화면에서도 하단에 배치된다.

기사와 정보원들의 가중치는 저널리즘 가치 중 사실성에 초점을 두고 저널리즘 관행에 따른 영역 지식과 NSNA을 통해 결정했다. 첫째, 정보원 가중치는 우선 개인실명정보원, 집단정보원, 익명정보원 순으로 할당된다. 개인실명정보원 간의 가중치 비교는 정보원의 논쟁적인 정도를 나타내는 값인 2주간 공동인용된 정보원수, 즉 연결정도 값을 활용한다. 집단정보원 간 가중치와 익명정보원 간 가중치는 동일하다. 둘째, 문장 가중치는 우선 개인실명정보원의 인용문, 수치문, 집단정보원의 인용문, 익명정보원의 인용문 순으로 부여된다. 인용문 내에서는 가중치가 높은 정보원의 인용문이 더 높은 가중치를 부여받는다. 셋째, 기사 가중치는 일자별 뉴스연결망분석에 의한 기사의 연결정도중앙성 값에 의해 부여된다. 이상의 것이 동일할 경우 대표 문장은 정보원, 수치 등 주요 사실 정보를 가장 많이 제시하는 문장으로, 대표 기사는 이러한 문장을 많이 포함하는 기사로 정했다.

정보원의 소속은 언론사의 출입처 관행과 지면 분류 관행을 참조해 대분류와 중분류 등 두 수준에서 분류했다. 활용된 분류표는 표 2와 같다.

표 2. 정보원 소속 분류표

대분류	중분류	계
정치	법원/검찰/변호사/경찰,소방/정당/국회/청와대/행정부/지자체/정부출연/정치인/북한/군대/국제기구	14
경제	농림수산/광업/에너지/IT전자/차,선박,항공/건설업/중소기업/석유화학/제철/대기업/패션,식품,마트/금융/서비스업/자영업/관광,외식/운송/협회/노조/직장인	19
종합/사회	의료계/학계/종교계/예술계/대중문화계/스포츠계	6
문화	시민봉사단체/대중교통/복지시설/시민/기타	5
국제	북미유럽/동아시아/중동, 아프리카/남미	4

기사의 지면은 우선 카인즈가 인력으로 분류한 지면 분류를 정보원 소속 분류에 맞추어 표 3과 같이 5개 분야로 재분류했다. 카인즈에서 분류하지 않은 지면에 대해서는 재분류된 기사를 참고해 각 분야에 대한 트레이닝 테이블을 작성한 후 간단한 베이지안 텍스트 분류(Naive Baysian Text classification)기법을 통해 추가로 분류했는데 평균 75%의 정확도를 기록했다. 여기서 미분류된 기사는 '종합/사회'에 할당했다.

표 3. 지면 분류표

뉴스소스 지면 분류	카인즈 지면 분류
정치	북한, 정치/해설
경제	정보통신/과학, 경제
종합/사회	지역, 사회, 생활, 여성, 종합. 특집, 인물/오피니언, 미분류
문화	방송연예, 스포츠, 매체, 문화
국제	국제, 외신

3) 시각화

(1) 메인화면

그림 4. 뉴스소스 메인화면 '오늘의 뉴스'

그림 4는 뉴스소스 어플리케이션 첫 화면인 '오늘의 뉴스' 화면이다. 단, 날짜 지정을 바꾸면 오늘로부터 과거 2주치 기사를 볼 수 있도록 디자인되어 있다. 기사는 5개 지면으로 분류되어 있으며 각 지면에서 가장 중요한 뉴스기사의 제목이 상단에 배치된다. 따라서 사용자는 모든 지면의 주요 기사 3~5개를 첫 화면에서 중복 없이 확인할 수 있다.

지면은 왼쪽부터 시작해 정치, 경제, 종합/사회, 문화, 국제 순으로 배치된다. 이는 지면의 가중치를 고려한 것이다. 기사 가중치를 부여하는 뉴스정보원연결망분석의 알고리즘 특성상 사실적이고 논쟁적인 기사가 더 중시된다. 이에 따라 지면별 가중치를 따져보면 대체로 정치, 경제, 종합/사회, 문화 순이 된다. 국제면의 경우 외국인 인명이 포함되어 자연어 처리 성능이 떨어지는 문제도 있다.

기본적으로는 각각의 대표 기사 제목이 보이는 칸은 모두 같

으며 매체명도 함께 보여준다. 다만 지면별 가중치 차이를 강조하기 위해, 가장 중요한 지면인 정치와 경제의 경우 최상단 기사는 기사 본문 내용의 일부를 함께 볼 수 있게 했다. 그 크기는 다른 지면 최상단 기사 제목의 두 배로 정했다. 또 기사별 가중치 차이를 강조하기 위해 1~3번째로 중요한 기사의 제목을 4번째 이하의 덜 중요한 기사 제목에 비해 두 배 크기로 보여준다. 스크롤을 통해 일단 20개의 대표 기사 제목을 볼 수 있으며, 추가로 '뉴스 더 보기' 메뉴를 선택하면 모든 대표 기사 제목들을 볼 수 있다.

사용자 경험 측면에서는 일반적인 뉴스포털의 사용자 경험과 일관성을 유지하는 한편 신문의 레이아웃도 참조해 디자인했다. 카인즈 아카이브 특성상 기사가 대부분 사진 없이 텍스트만 제공되는데다가, 첫 화면은 모두 제목으로 이뤄져 있다. 신문의 레이아웃은 텍스트 중심인 기사를 배열할 때 느끼는 시각적 단조로움을 줄이기 위해 다양한 방식을 시도한다. 뉴스소스에서는 가중치에 따라 제목의 글자 크기와 굵기, 제목 간 구분 시 실선 및 점선의 사용 등을 활용했다.

'오늘의 뉴스' 화면에서 사용자가 대표 기사 제목을 선택하면 그림 5처럼 해당 기사의 원문과 함께 하단에 대표 기사 및 유사 기사에 포함된 사실정보 군집들이 중요도 순으로 한 화면에 보인다. 뉴스소스에서 보이는 사실정보들은 정보원의 인용문과 수치 등이며 기사 본문 속 문장 형태 그대로 제시된다. 정보원의 경우 이름・소속・직함 외에도 해당 정보원의 전체 인용건수, 해당 정보원의 인용문 중 가중치가 가장 높은 인용문 일부가 제시된다. 사실정보 군집들이 많을 경우 가장 중요한 사실정보 군집 3건이 우선 제시된다.

그림 5. 뉴스제목 선택화면

'정보 더 보기'를 선택하면 그림 6과 같이 대표 기사와 관련된 모든 사실정보를 볼 수 있다. 이 때 사실정보가 여러 매체에 중복되어 있을 경우 중복을 제거하고 대표 정보만 보인다. 이 기능으로 사용자는 같은 내용을 다루는 다양한 매체의 기사를 중복해 볼 필요 없이 다양한 사람들의 의견과 관련 사실을 중요한 순서대로 한 화면에서 살펴볼 수 있다.

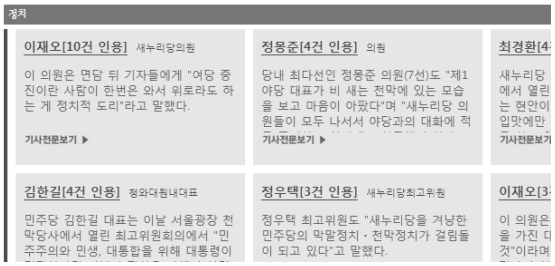

그림 6. 추가정보원이 제공된 화면(좌변 확대)

'기사 전문 보기'를 선택하면 그림 7과 같이 별도의 창이 떠서 해당 사실정보가 포함된 기사를 볼 수 있다. 사실정보 군집이 정보원의 인용문일 경우 화면을 반으로 나누어 왼쪽에는 해당 정보원의 인용문들이 가중치 순으로 제시되며, 화면의 오른쪽에는 각 인용문들이 포함된 기사 본문이 나타난다. 이 때 인용문은 하이라이트된다.

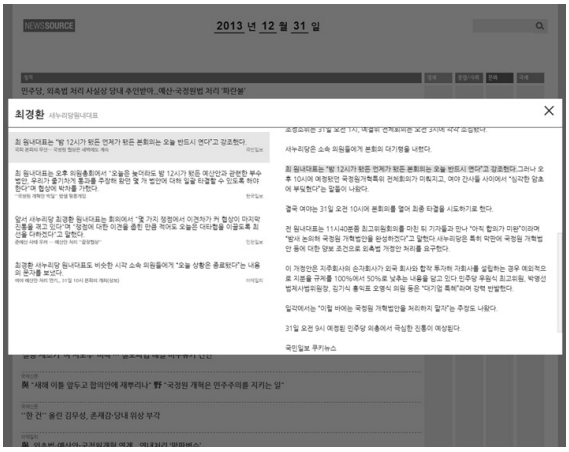

그림 7. 기사 보기 화면

(2) 검색화면

뉴스소스의 또 다른 기능으로 주제어 검색 기능이 있다. 이 기능은 기본 지정된 기간 또는 사용자가 설정한 기간 동안 검색어가 포함된 모든 뉴스 기사를 대상으로 분야 및 소속별 정보원들을 중요한 순서대로 제공한다. 이 기능을 통해 사용자는 관심 있는 주제에 관해 최소한 언론의 관점에서 중요하는 정보원이 누구이고 그들의 의견은 무엇인지 소속별로 확인할 수 있다. 검색화면을 통해 특정 주제어에 대한 내용을 다양하고 심층적으로 비교할 수 있는 셈이다.

일반적인 뉴스포털와 비교하면 기사 검색 시 기사에 포함된 정보원이 누구인지는 확인할 수 있지만, 그 정보원이 다른 정보원에 비해 얼마나 더 중요한지, 심지어 중요한 정보원이기는 한

지 파악하기 힘들다. 설사 단순 인용빈도를 파악한다하더라도 연예인처럼 중복기사가 양산되는 정보원의 중요성이 과대평가 될 수 있다. 반면 뉴스소스의 검색 기능을 통해서 사용자는 우선 원하는 주제에 대해 최소한 언론사가 중시하는 권위 있는 정보원이 누구인지를 세세하고 쉽게 파악할 수 있다. 예컨대 2012년 상반기에 '금융위기'를 검색어로 입력했다고 치자. 기존 뉴스포털에서는 이명박 대통령이나 박재완 기획재정부 장관 등이 중요한 정보원이라는 것은 쉽게 예상할 수 있지만, 예컨대 전 금융위원회 상임위원인 이종구 변호사나 거시경제전문가 유병규 현대경제연구원 경제연구본부장 등이 이들과 버금가게 중요한 것은 해당 분야 전문가가 아니라면 쉽게 알 수 없고, 이들의 발언을 추려서 검토하는 것은 더더욱 어렵다. 반면 뉴스소스 검색기능을 활용하면 금융위기에 대해 잘 모르는 일반인이라고 할지라도 언론사가 가장 중시하는 이러한 정보원의 명단과 전문적 의견을 쉽게 파악할 수 있다.

그림 8은 기간을 2013년 9월 1~5일로 설정한 뒤 '창조경제'라는 주제어를 검색한 결과이다. '창조경제'라는 주제어에 관해 정보원들이 소속 기준으로 대분류(정치, 경제, 종합/사회, 문화, 국제)에 따라 배열된다. 그림 8의 경우 국제 분야는 검색된 인물이 없으므로 페이지 공간 활용을 위해 영역이 줄어들어 있다. 이렇게 배열하면 우선 해당 주제어에 대해 소속의 대분류별로 최상위권의 중요도를 갖는 정보원의 의견을 한 눈에 파악할 수 있다. 각각의 정보원들은 대분류에 따라서는 동일한 계통의 색을 갖는다. 그러나 중분류에 따라서는 서로 다른 색상을 가진다. 예컨대 그림 7에서 정부 소속의 '윤상직'과 변호사인 '배정환'은 같은 계통의 색을 가지나 소속이 다르기 때문에 다른 색상의 선으로 구분된다.

그림 8. 주제어 '창조경제'의 검색결과(좌변 확대)

한편 정보원들을 단순히 중요도 순서에 따라 배열할 경우 특정 소속의 입장만 집중되는 문제가 발생할 수 있다. 이를 방지하고 다양성을 확보하기 위해 뉴스소스에서는 우선 표 2의 중분류 소속별로 가장 중요한 정보원을 먼저 배열하고, 일순한 뒤에는 각 소속별로 두 번째로 중요한 정보원을 배열한다. 예컨대 검색 결과 정보원들이 정치 분야에서 5명의 여당 소속 정보원과 2명의 야당의 소속 정보원으로 나오고, 여당의 인원들이 모두 중요도 1~5위를 기록하고 야당의 인원이 6~7위를 기록한다고 치자. 그럴 경우 단순 가중치 순서로 배열하게 되면 여당 소속 정보원이 상위 1~5위에 배열된 다음 야당 소속 정보원이 나와서, 첫 화면에서는 여당 정보원 의견만 보게 되고 야당 정보원 의견은 목록의 하단까지 내려야 확인할 수 있는 문제가 발생한다. 하지만 뉴스소스에서는 소속별 1순위 정보원을 먼저 보여주기 때문에, '여당 1위-야당 1위-여당 2위-야당 2위-여당 3위-여당 4위-여당 5위' 순으로 재배열된다. 다음으로 각 정보원의 '기사 보기' 기능은 앞서 메인화면과 동일하게 정보원을 클릭하면 새로운 창이 뜨고 창의 왼쪽에는 정보원의 인용문이, 오른쪽에는 기사 본문이 제시된다.

검색화면에서는 간단한 자아연결망도 시각화한다. 뉴스정보원연결망에서 자아연결망은 자아에 해당하는 한 정보원이 자신과 같은 기사에 함께 인용된 정보원과 만드는 연결망을 뜻한다. 이는 특정 주제에 대해 선택한 자아 정보원의 의견에 직접적으로 반박하거나 뒷받침해주는 정보원들의 의견들을 보여준다. 그림 8은 그림 7과 마찬가지로 주제어를 '창조경제'로 하고 같은 기간으로 검색한 뒤, 박근혜를 선택하고 자아연결망 기능을 활성화한 것이다. 우측 상단에는 '박근혜'와 함께 인용된 정보원이라는 설명이 뜨며, 선택된 자아 정보원인 '박근혜'의 자아연결망에 포함되지 않는 정보원의 경우 사라지고, 화면에는 자아연결망에 포함되는 정보원만 남아있게 된다. 실제로 '박근혜'의 자아연결망에 포함되는 정보원으로 경제 분야의 '우원길'과 종합/사회 분야의 '최문기' 등이, 문화 분야에서는 '이경재' 등이 남아있는 것을 확인할 수 있다. 이들은 "창조경제"라는 주제와 관련해 '박근혜'와 함께 기사에 인용된 정보원들이다.

그림 9. 주제어 '창조경제'의 '박근혜'의 자아연결망

4. 결론

온라인매체의 등장으로 뉴스 기사 양은 급격하게 늘었다. 하지만 흥미 위주의 연예기사가 대다수를 차지하는데다가 그나마도 중복된 기사가 너무 많다. 또한 사실과 관련 없이 논평 기

사가 양산되면서 사회적 갈등도 심화된다. 게다가 기존의 뉴스 포털서비스가 인기 위주로 가중치를 부여하면서 이러한 현상을 더욱 부채질한다. 결국 사회문제를 사실에 입각해 논쟁적으로 다루는 기사들이 지나치게 묻히게 된다.

본 논문은 저널리즘의 가치 중 사실성, 비판성, 다양성, 심층성 등이 높은 기사와 정보원에 가중치를 두는 NSNA을 기반으로 일반 사용자들이 쉽게 이해할 수 있는 뉴스소스를 제안하고 실제로 구현한 웹어플리케이션을 소개했다.

뉴스소스는 유사한 중복 기사나 문장은 대표 기사와 대표 문장으로 축약하고, 축약된 기사들 중에서 중복되지 않은 사실정보들을 추려서 함께 제시한다. 사용자는 가중치에 따라 변형된 누적막대그래프 형태로 제시되는 기사를 통해 중복된 정보를 필터링하고 최대한 다양한 분야의 의견을 한 화면에서 볼 수 있다. 더 나아가 뉴스소스는 단순한 키워드가 아니라 문장 단위의 내용을 중복 없이 다양한 소속의 관점에서 중요도 순으로 검토할 수 있게 해준다. 즉 검색결과가 일종의 약식 보고서처럼 제시되는 것이다.

요컨대 뉴스소스는 기사를 분석하는 언론학자나, 관련 기사를 검토하는 기자, 기업 홍보팀, 정부 공보팀 등 언론 관계자는 물론, 새로운 프로젝트에 착수하는 변호사나 컨설턴트, 신규 시장 진입을 검토하는 마케터, 공청회를 준비하거나 자문위원을 꾸려야하는 정부 관계자 등에게 도움을 줄 수 있다. 수업 과제물을 작성하는 대학생, 사회적 쟁점을 다각도로 검토하고 싶은 뉴스 중이용자 등 일반인에게도 훨씬 깊이 있는 뉴스 어그리게이션 서비스(news aggregation service)로 활용될 가능성을 갖는다.

다만 본격적인 서비스를 위해서는 자연어처리 성능을 개선할 필요가 있다. 기존 뉴스포털과 다소 중복되기는 하지만 그럼에도 불구하고 기사 검색 등 여러 기능을 추가할 필요도 있다. 언론사와의 저작권 문제도 남아있다. 텍스트 중심에서 사진이나 동영상을 함께 서비스할 때 시각화 방안도 고민할 필요가 있다. 또한 본 연구의 기초가 되는 NSNA를 더욱 정교화해 기사와 사실의 가중치 부여 방식을 다양화하고 이를 바탕으로 준공공데이터인 기사로부터 더 많은 가치를 끌어낼 수도 있을 것이다.

참고문헌

[1] 한국언론연구원. 매스컴대사전. 서울 : 한국언론연구원. 1993.

[2] 이준웅. 인터넷 공론장의 매개된 상호가시성과 담론 공중의 형성. 언론정보연구. 제46권 2호. pp. 5-32. 2009.

[3] 김위근, 김성해, 김동윤. 뉴스의 대중화 혹은 저널리즘의 게토화 : 저널리즘 관점에서 본 네이버 '뉴스캐스트' 사례 분석. 사이버커뮤니케이션학보. 제30권 2호. pp. 33-72. 2013.

[4] van Dijk, T. A. News As Discourse. Lawrence Erlbaum. New Jersey. 1988.

[5] Sigal, L. V. Reporters and Officials : The Organization and Politics of Newsmaking. Lexington. DC Health & Co. Lexington. Massachusetts. 1973.

[6] Schudson, M. Discovering The News : A Social History of American Newspapers. Basic Books. New York. 1978.

[7] Gans, H. Deciding Whats News. University of Texas Press. Austin. 1979.

[8] 박대민. 뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망분석. 한국언론학보. 제57권 6호. pp. 233-261. 2013.

[9] Weskamp, M. Newsmap. Webdesigning Magazine. June 2004.

[10] Bederson, B. B., Shneiderman, B., and Wattenberg, M.. Ordered and Quantum Treemaps : Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics. 21(4). pp. 833-854. 2002.

[11] http : //media1.s-nbcnews.com/i/msnbc/components /spectra/ 2013.

[12] Witten, I. H., Frank, E. and Hall, M. Data Mining : Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann. 2011.

박 대 민

1998년 3월 ~ 2003년 8월 서울대학교 언론정보학과 졸업(문학학사). 2003년 9월 ~ 2006년 8월 서울대학교 대학원 석사 졸업(문학석사). 2006년 9월 ~ 2014년 8월 서울대학교 대학원 졸업 (문학박사). 2014년 8월~현재 한국언론진흥재단 선임연구위원.

김 기 남

2007년 3월 ~ 2013년 2월 경기대학교 컴퓨터과학과 졸업(이학사). 2013년 3월 ~ 현재 아주대학교 대학원 라이프미디어협동과정. 관심분야는 데이터마이닝, 웹사이언스, 비주얼라이제이션임.

강 남 용

2011년 3월 ~ 2014년 2월 아주 대학교 컴퓨터공학과 졸업(공학사). 2014년 3월 ~ 현재 한국과학기술원 대학원 석사과정. 관심분야는 사물 인터넷, 사물 간 신뢰도 측정임.

서 봉 원

1989년 3월 ~ 1993년 2월 서울대학교 계산통계학과 졸업(이학사). 1993년 3월 ~ 1995년 2월 서울대학교 대학원 계산통계학과 졸업(이학석사). 1998년 8월 ~ 2005년 5월 Univ. of Maryland at College Park 대학교 대학원 졸업 (전산학박사). 2013년 9월~현재 서울대학교 융합과학부 교수.

하 효 지

2010년 3월 ~ 2013년 8월 아주 대학교 미디어학과 졸업(미디어학사). 2013년 9월 ~ 현재 아주 대학교 대학원 석사과정. 관심분야는 정보시각화디자인, 인간-컴퓨터 상호작용, UX/UI 디자인임.

온 병 원

1991년 3월 ~ 1998년 2월 안양대학교 컴퓨터공학과 졸업(공학사). 1998년 9월 ~ 2000년 8월 고려대학교 대학원 컴퓨터학과 졸업(이학석사). 2002년 9월 ~ 2007년 8월 펜실베이니아주립대학교 대학원 졸업 (공학박사). 2014년 4월 ~ 현재 군산대학교 통계컴퓨터과학과 조교수.