N3WS : 키워드 및 요약문장 추출을 이용한 인터랙티브 신문기사 탐색

조희정*, 손지연*, 윤별이*, 조아현*, 김 명*, 박은정**
*이화여자대학교 컴퓨터공학과
**네이버 파파고
heybb816@ewhain.net, gyouns@ewhain.net, stardl@ewhain.net, ccho0513@ewhain.net, mkim@ewha.ac.kr, lucy.park@navercorp.com

N3WS : Interactive Newspaper Article Navigation Using Keyword and Summary Extraction

Hee-Jeong Cho*, Ji-Youn Son*, Byeol-Yi Yoon* , A-Hyun Cho*, Myung Kim*, Eun-Jeong Park**
*Dept. of Computer Science & Engineering, Ewha Womans University
**Dept. of Papago, Naver Corp.

요 약

최근 인터넷 기사 중에는 부정확한 제목이나 자극적인 단어를 사용하는 경우가 많아 구독자에게 불편함을 준다. 본 논문에서는 이러한 기사들의 헤드라인을 삭제하고, 기사의 내용을 3 문장으로 요약해 주어, 구독자가 원하는 기사를 효율적으로 파악할 수 있게 하는 시스템을 제안한다. 제안하는 본 시스템은 파이썬 언어의 KoNLPy 패키지를 사용하여 기사의 단어들을 형태소 단위로 분석하며, 추출된 키워드를 토대로 워드 클라우드를 생성한다. 사용자가 클라우드의 특정 단어를 선택하면, 해당 신문기사들의 본문을 분석하여 각 신문 기사만의 핵심적인 문장을 3 문장으로 출력해 준다.

1. 서론

언론 매체들의 수익은 구독자 수에 비례한다. 구독률을 높이기 위해, 일부 매체들은 본문의 내용과 상관 관계가 없는 자극적이고 선정적인 기사 제목으로 구독자를 유인하곤 한다. 예를 들어, 2017 년 1 월~3 월 사이에 신문 윤리 위원회의 제재를 받은 기사 중에는 낚시성 제목을 갖는 기사가 4 분의 1 이나 된다는 연합뉴스의 기사 [1]가 있었다. 낚시성을 띠지 않는 헤드라인이라고 해도, 제목만을 보았을 때 해당 기사의 내용을 쉽게 유추하기 어려운 기사들도 다수 존재한다. 독자들을 속이거나 정확한 사실을 전달하지 않는 기사 제목으로 인해 구독자들은 많은 불편함을 느끼게 된다. 이는 또한 해당 언론매체의 신뢰성 하락으로 이어질 수 있다.

현대 사회는 다양한 기술의 급속한 발전에 따라 '더 빠르고, 더 쉬운 것'을 요구한다. 종이 신문을 읽는 것보다 스마트폰으로 인터넷 뉴스를 보는 현대인이 더 많다. 또한 긴 글의 본문을 읽기보다는 길이가 짧은 헤드라인만을 읽는 경향이 크다. 이러한 현상이 더욱 심화되고 있는 현 사태에서 신문 기사의 제목은 큰 중요성을 띠고 있으며, 현재 제시된 문제들은 반드시 보완되어야 할 점들이다.

해당 문제점들을 해결하려면, 언론에서는 독자들이 원하는 기사를 제대로 읽을 수 있는 환경을 제공해야 하며 빠른 시간 내에 기사 본문의 내용을 정확하게 파악할 수 있게 하는 방법을 제시해 주어야 한다. 본 논문은 이러한 문제점을 해결하는 방안으로, 혼란을 일으킬 수 있는 헤드라인을 삭제하고 각 기사의 본문을 3 줄로 요약하여 원하는 기사를 효율적으로 찾아 읽을 수 있는 시스템을 제안한다.

본 논문은 다음과 같이 구성된다: 2 절에서 자연어 처리를 사용한 기존의 문서 요약 사이트들을 살펴보고, 3 절과 4 절에서는 본 논문에서 제안하는 N3WS 시스템에 관해 기술한다. 끝으로 5 절에서 본 논문의 결론을 맺는다.

2. 기존 연구 사례 분석

2.1 국외 사례

텍스트 요약(text summarization)에 대한 다양한 선행 연구가 수행되어 있고 [2] [3] [4], 하나의 긴 영문 글을 원하는 길이로 요약해주는 웹 서비스 [5] [6]도 구현되어 있다. 'Smmry' [5]의 경우 영문으로 된 글을 텍스트 또는 PDF 파일 형태로 입력받아 사용자가 지정하는 문장의 수로 해당 글을 요약해 준다. 이와 유사한 프로그램으로 'Text compactor'라는 프로그램 [6]도 있다.

2.2 국내 사례

국내 사례로는 국민대학교 한글 공학 정보 검색 연구소에서 개발한 한국어 처리 시스템인 'Korean Language Processing System' [7]이 있다. 해당 프로그램은 형태소 분석, 명사 추출, 자동 문서 분류 등 한국어를 이용한 형태소 분석, 합성 등 한국어의 맥락을 이해하고 내용을 분류해 주는 기능이 있다. 본 연구 프로젝트와 가장 큰 유사성을 갖는 프로그램인 '세줄 요약기' [8]는 장문의 글을 3 문장으로 요약해주는 사이트이다. 그림 1 은 '세줄 요약기'가 글을 요약해 주는 예제이다.

(그림 1) 세줄 요약기의 요약 사례.

한글은 영어와는 달리 명사의 복수 표현, 동사의 과거, 현재, 미래 표현 등이 복잡하기 때문에 분석하기가 쉽지 않다. 따라서 영문 요약 시스템들과는 달리 한글 요약 프로그램들은 찾아보기 힘들다. 본 연구에서 제안하는 시스템은 국내의 유사 사례들과 달리 사용자의 사용 시간에 따라 자동으로 여러 개의 글을 크롤링하여 수집하고, 분석한 후 사용자가 원하는 글들을 요약해서 보여준다는 면에서 차별점을 갖는다.

3. N3WS : 인터렉티브 신문기사 탐색 시스템

본 프로젝트에서 제안하는 N3WS 시스템은 그림 2 와 같은 구조를 갖는다. 이 시스템은 사용자의 접속 시간을 기준으로 신문 기사를 크롤링하여 워드 클라우드로 보여주는 부분과 사용자가 선택하는 키워드에 관한 기사들의 세줄 요약을 보여주는 부분으로 구성된다. 이제 각 부분에 대해 구체적으로 설명하고자 한다.

(그림 2) N3WS 시스템 구성도.

3.1 기사 크롤링 및 워드 클라우드 생성

사용자가 N3WS 웹사이트에 접속하면, 그 접속 시간을 기준으로 최근 기사들을 크롤링한다. 초기 데이터베이스로 약 2,000 개의 기사를 크롤링 하며, 10 분 간격으로 새로 업로드 된 기사를 검색하여 자동으로 저장한다.

크롤링 된 기사들에서 키워드를 추출한 후, 해당 기사의 주소와 본문을 함께 뉴스 정보를 저장하는 데이터베이스(이하 NEWS DB)에 저장한다. 이 때, 전체 기사를 대상으로 빈도수가 높은 키워드를 추출하였을 때 보다, 각 기사당 높은 빈도수를 가진 3 개의 키워드를 추출하고 그 빈도수를 다시 계산하는 것이 정확도가 높았다. 따라서, NEWS DB 에는 키워드 1, 키워드 2, 키워드 3 을 저장하며 해당 단어들을 워드 클라우드에 사용하도록 설정한다. "이대", "이화여대"와 같이, 동일한 대상을 지칭하는 단어들은 딕셔너리를 사용한 함수를 이용하여 미리 통일시키며, 동일한 함수를 통해 "일보", "것" 등 키워드로 인식하지 않을 단어들을 사전에 제거한다. 그림 3 은 초기 NEWS

DB 의 예시이다. 저장된 키워드들과 그 빈도수를 딕셔너리 형식으로 저장한 후, 그 내용을 CSV 파일에 저장하여 이를 토대로 워드 클라우드를 생성한다.

(그림 3) 초기 NEWS DB.

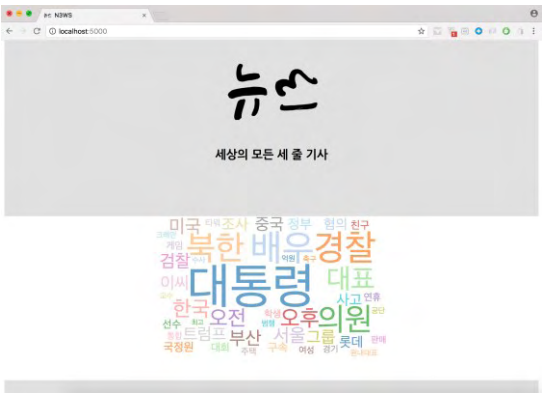
3.2 특정 키워드에 대한 3 문장 요약

워드 클라우드에서 사용자가 키워드를 선택하는 순간, 해당 키워드를 포함하고 있는 기사들을 최대 20 개까지 데이터베이스에서 로드한다. 해당 기사들은 워드클라우드 생성 때와 동일한 전처리 과정을 거친 후 명사, 동사, 형용사의 형태소를 각 기사별 리스트로 저장한다. 모든 기사의 형태소 분석이 끝나면 각 형태소 리스트를 대상으로 TF-IDF 계산을 통한 키워드 추출이 실행되며, 기사 별로 중요한 형태소 3 개를 다시 추출한다.

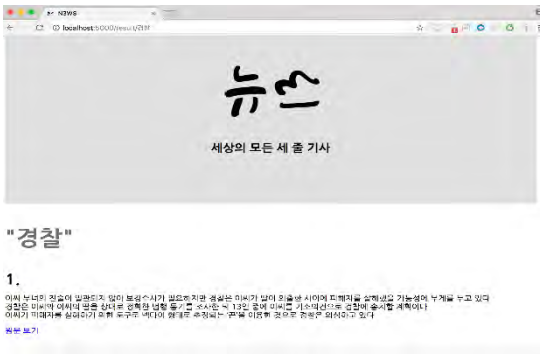
이 때 추출된 키워드는 타 기사에 비해 해당 기사에서만 두드러지게 나타나는 형태소들이다. 이러한 형태소들을 포함하고 있는 문장 중, personalized PageRank 를 통해 중요도가 높은 문장을 계산해내어 그 중 상위 3 개를 출력하는 것이 본 알고리즘의 최종 결과물이다. 이때 세 문장은 포함하고 있는 형태소의 특성상 처음에 선택한 주제 키워드에 관해 각 기사만이 가진 내용을 요약한다. 따라서 특정 주제에 대한 여러 기사들의 효율적인 요약문을 제시한다.

3.3 웹사이트 구현

N3WS 는 그림 4 와 같이 구현되어 있다. 초기 화면의 상단에는 로고가 자리 잡고 있으며 로고 하단부에는 워드 클라우드가 자리잡고 있다. 반응형 웹 사이트 구현을 위해 어느 사이즈의 화면 크기이든지 중앙에 위치하여 사용자의 편리성을 도모하였다.



(a) 홈 페이지



(b) 키워드 선택 페이지

(그림 4) 웹사이트 구현 화면.

워드 클라우드의 키워드 중에서 사용자가 원하는 키워드를 선택하여 클릭하게 되면 기사별로 주요 문장을 출력하는 두 번째 페이지로 연결된다. 키워드를 선택한 시점에서 해당 키워드에 대한 기사를 로드하여 계산하기 때문에 기사의 개수에 따라 잠시 시간이 지연될 수 있다. 이러한 불편함을 줄이기 위해 NEWS DB 의 업데이트와 동시에 업데이트 되는 캐시 데이터베이스(이하 CACHE DB)를 설정하였다. 선택한 키워드가 처음 선택된 단어일 때, CACHE DB 에 해당 키워드, 원문 주소, 요약 문장들을 저장하며 후에 같은 키워드가 선택되었을 때 별도의 계산 과정을 거치지 않고 CACHE DB 에서 데이터를 로드할 수 있도록 설정하였다. 그림 5 는 CACHE DB 의 예시이다.

두 번째 페이지로 이동하면 좌측 상단 부분에 사용자가 선택한 키워드를 두꺼운 글씨로 출력한다. 그리고 그 키워드 아래에 기사별로 3 줄 요약문을 최신 순으로 게시한다. 요약 문장들 아래에는 기사의 원문으로 연결되는 글귀를 두어 사용자가 자유롭게 해당 기사의 원문 사이트로 이동할 수 있도록 하였다. 또한, 우측 하단에 페이지의 상단으로 다시 올라갈 수 있는 장치를 넣어 사용자에게 편리함을 제공하였다.

http://news.donga.com/List/3/all/20171011/86701518/1	김정문 NC 다이노스 감독, ⓒ News1 김정문 NC 다이노스 감독이 '원점 승부'라는 자세로 준플레이오프 3차전을 맞는다. 김 감독은 11일 창원 마산구장에서 열리는 2017 타이어뱅크 KBO리그 롯데 자이언츠와 준플레이오프 3차전을 앞두고 "1승1패인 원점이다. 지난 경기는 잊고 오늘 이기면 유리하다는 생각으로 임할 것이라고 말했다. NC는 SK 와이번스와 와일드카드 결정전을 승리한 뒤 준플레이오프 1차전에서도 9-2로 이겼다. 그러나 2차전에서 빈공에 허덕이며 0-1로 패배 1승1패 동률이 됐다. 3차전을 내줬기 때문에 기세는 롯데 쪽이 낫다는 평가. 3차전이 큰 부담으로 작용할 수 있는 상황이지만 김 감독은 그저 '원점에서 다시 시작한다'는 생각으로 경기를 준비하고 있다. 이번 준플레이오프는 두 팀 모두 공격력이 시원치 않다. NC가 1차전서 9-2로 이겼지만 연장전에서 대거 7점을 냈을 뿐 점수이닝까지는 2득점에 그쳤다. 김 감독은 "오늘은 롯데 타격 컨디션이 좀 나아질 것이다. 우리도 집중해서 점수를 내야 한다"며 "12차전이 낮경기로 열렸다는 점도 고려해야 한다. 낮경기에는 타자들이 빠른 공에 잘 적응이 늦다"고 조심스럽게 타격전을 예상했다. 비록 패했지만 2차전에서 호투한 장현식에 대한 칭찬도 잊지 않았다. 장현식은 7이닝 동안 비자책 1실점으로 롯데 타선을 잠재웠으나 타선의 도움을 받지 못하고 패전투수로 기록됐다. 김 감독은 "6이닝만 던져줘도 대단한 것인데 7이닝을 던졌으니 정말 대단한 것"이라며 "(장)현식이에게 승운이 따르지 않는 경우가 많은데 형들한테 좀 처달라는 뜻에서 좋은 것 좀 돌리라고 농담을 했다"고 전했다. (청원=뉴스1)	감독	플레이오프	롯데
http://news.donga.com/List/3/all/20171011/86701618/1	코스피가 전 거래일보다 1.00%(24.35포인트)오른 2458.16에 장을 마감하며 장중 -종가 기준 최고치를 일제히 경신했다. 11일 서울 중구 KEB하나은행 외환 딜링룸. 홍진환 기자 jean@donga.com 코스피가 글로벌 증시 강세와 돌아온 외국인에 힘입어 사상 최고치를 돌파했다. 11일 코스피는 전날보다 1.00% 오른 2458.16에 거래를 마쳤다. 지난 7일 24일 세운 종가 사상 최고치 기록(2451.53)을 두 달 만 만에 갈아 치운 것이다. 간밤 뉴욕증시가 상승 마감하면서 코스피도 상승장을 연출했다. 10일(현지시간) 뉴욕증시는 3분기 실적 기대감에 다우존스산업평균지수와 나스닥 지수가 나란히 사상 최고치를 경신했다. 대장주 삼성전자는 전날보다 3.48% 오른 273만2000원에 거래를 마치며 사상 최고가를 경신했다. 삼성전자는 13일 3분기 잠정실적을 발표한다. 글로벌 반도체 가격 상승으로 역대 최대 실적을 낼 것으로 증권업계는 전망하고 있다. 신민기 기자 minki@donga.com	최고	코스피	사상
http://news.donga.com/List/3/all/20171011/86701639/1	- 리니지M 100일간의 기적 프로모션 실시 *엔씨소프트(대표 김택진)는 자사에서 개발 및 서비스 중인 모바일 MMORPG 리니지M에 10월 11일부터 24일까지 천만 명이 만들어낸 100일간의 기적 프로모션을 진행한다.* "리니지M 이용자는 게임 내 상점에서 100 아데나(게임 내 재화)로 T.I의 100일 기념품을 구입할 수 있다. T.I의 100일 기념품을 사용하면 T.I의 서신 아이템과 100일 기념 반지 아이템을 획득할 수 있다." "T.I의 서신 아이템은 희귀 방어구 제작 비법서를 대체할 수 있는 아이템이다. 100일 기념 반지 아이템을 착용하면 이벤트 기간 동안 방어력이 높아진다. 100일 기념 반지 아이템은 이후 1주년 기념 반지 제작 재료 아이템으로 활용될 예정이	게임	아이템	블리즈컨

(그림 5) CACHE DB.

4. 결론

본 논문에서는 자연어 처리를 통해 인터넷 기사의 제목 대신 차별화 된 핵심적인 내용을 3 줄 문장으로 재구성하는 방법을 제안하였다. 신문사의 최신 기사들로부터 워드 클라우드를 생성하며, 워드 클라우드에서 원하는 키워드를 누르면 그와 관련된 기사들을 나열해 준다. 이때, 기사들은 원래의 제목은 제거하고, 3 줄로 된 요약을 제공해 준다.

본 연구에서 제안한 시스템은 사용자의 요구에 따라 인터랙티브하게 신문기사 탐색을 가능하게 한다. 따라서 단시간 내에 효율적인 정보 습득을 제공하려는 다양한 분야에서 활용이 가능 할 것으로 기대한다.

국외와 국내에 본 논문에서 제시한 방법과 유사한 사례들이 있지만, 국내에는 아직 한국어를 대상으로 하는 자연어 처리 기술이 많이 미약하다. 본 논문에서 제안하는 방식으로 분석을 한다면 한국어 자연어 처리에 새로운 방법을 제안할 수 있다고 예상한다.

ACKNOWLEDGMENT

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 서울어코드활성화지원사업(IITP-2017-2012-0-00442)의 연구결과로 수행되었음

참고 문헌

[1] aupfe@yna.co.kr, "연합뉴스," 14 04 2017. http://www.yonhapnews.co.kr/bulletin/2017/04/13/02 00000000AKR20170413152000033.HTML.

[2] F. R. Mohamed Abdel Fattah, "Automatic Text Summarization," %1 World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:1, 2008.

[3] H. Jing, "Sentence Reduction for Automatic Text Summarization".

[4] 김. 김. 차준석, "단어 간 의미적 연관성을 고려한 어휘 체인 기반의 개선된 자동 문서요약 방법," 2017.

[5] http://smmry.com/.

[6] http://textcompactor.com/.

[7] http://nlp.kookmin.ac.kr/demo/index.html.

[8] https://summariz3.herokuapp.com.

[9] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지," 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014.

[10] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출," 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제 36 권 제 1 pp. 311~314, 2009.

[11] 설진석, 이상구, "lexrankr: LexRank 기반 한국어 다중 문서 요약," 한국정보과학회 2016 동계학술발표회 논문집, Vol. 2016, No. 12, pp. 458~460, 2016.