논문 18-43-04-11 The Journal of Korean Institute of Communications and Information Sciences '18-04 Vol.43 No.04
https://doi.org/10.7840/kics.2018.43.4.693

인터넷 정보의 초기 특성 및 분야별 다양성 분석

노 기 섭◆, 오 하 영°

Analysis on Initial Characteristics and Field Diversity of Internet Information

Giseop Noh◆, Hayoung Oh°

요 약

최근 정보 생성 및 제공 방식은 ICT 기술의 발전으로 급격히 변화하고 있다. 이러한 정보화의 발전은 정보 소비자들에게 순기능으로 작용하기도 하지만, 무분별한 개인정보 유출, 온라인 따돌림, 가짜뉴스 유통 등의 부작용이 발생하기도 한다. 그러나 정보화 부작용을 분석하기 위한 정보 생애 관점의 접근은 부족하였다. 본 논문에서는 인터넷 정보 생애를 '생성(birth) - 성장(growth) - 침체(stagnation) - 소멸(death)' 4단계로 정의하고 정보의 성장 단계 분석 방법을 제시한다. 본 논문은 인터넷 정보 성장을 측정하기 위한 4가지 측정 지표(긍정도, 토론수준, 참여도, 민감도)를 새롭게 제시하고, 자동화된 데이터 크롤러(crawler)를 구현하여 실제 온라인 데이터를 수집한다. 온라인 데이터는 카테고리, 키워드별로 연관 기사를 수집하여 분석한 결과 주제별로 또는 키워드별 온라인 정보 성장 수준이 상이함을 확인하였다.

키워드 : 정보생애, 정보 성장, 성장 레벨, 참여도, 민감도

Key Words : Information Life, Information Growth, Growth Level, Attendance, Sensitivity

ABSTRACT

Recently, the means of the generation of information and provision has been changed rapidly as the ICT technologies grow up drastically. While the development of this information is acting as a pure function to information consumers, there are side effects such as irrelevant personal information leakage, online bullying, and fake news circulation. In this paper, we define the Internet information life as four stages of 'birth - growth - stagnation - death' and present a method of analyzing the growth stage of information. In this paper, we present four new metrics (optimistic level, discussion level, attendance level, sensitivity) to measure Internet information growth and collect actual online data by implementing an automated data crawler. As a result of analyzing the related articles by category and keyword, online data shows that the level of online information growth by topic or keyword is different.

※ This work was supported by the Ajou University research fund and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1A1B03035557).

◆ First Author : Cheongju University, Division of Software Convergence, kafa46@gmail.com, 정회원

° Corresponding Author : (ORCID:0000-0002-7362-5138)Ajou University, DASAN Colleage University, hyoh79@gmail.com, 정회원

논문번호 : KICS2018-01-014, Received January 10, 2018; Revised January 31, 2018; Accepted February 5, 2018

Ⅰ. 서 론

최근 정보생성 및 제공 방식은 ICT 기술의 발전으로 급격히 변화하고 있다. 현대의 정보생성 및 제공은 과거의 종이 형태의 신문, 잡지, 소식지 등에서 인터넷 뉴스, 온라인 사회 관계망 서비스(online social network service) 등으로 급격히 이동하고 있다. 온라인을 활용한 정보제공 수단의 변화는 정보 소비자들이 신속하게 해당 정보를 수용 및 소비하고 새로운 여론을 신속하게 형성하여 국가 정책에 반영되거나 사회적 순기능 역할을 담당하기도 한다.

그러나 온라인 정보제공은 사생활 정보의 무분별한 유통, 온라인 따돌림[1], 가짜(fake) 뉴스[2]의 빠른 확산 등 사회적 악영향도 발생시키는 등 순기능과 악기능의 양면을 동시에 가지고 있다. 심각한 경우 가짜 뉴스는 언론 전체의 신뢰성을 저하시키고 온라인 따돌림은 피해자를 자살에 이르게 한다.

인터넷에서 생성되고 유포되는 정보들의 중요한 특징은 빠른 확산 속도이며, 한번 전파된 정보는 추가적인 정보 유통을 차단하기도 어렵고 온라인 공간에서 완전히 삭제하기도 어렵다. 따라서 온라인 정보의 초기 상태를 분석하여 적절한 대응 방안을 강구하는 것이 중요하다. 정보전달 방법, 최대화, 최소화 등에 대한 연구는 진행되었으나, 온라인 정보의 초기 생성과정을 분석하고 특성을 파악하는 것에 대한 연구는 부족하였다.

인터넷 네트워크 또는 온라인 소셜 네트워크 분석을 이용하는 정보생성과 확산형태에 대한 연구가 있었지만 초기정보 형태를 분석하여 부적절한 정보를 차단하고 올바를 정보를 파악하여 정보확산을 증가시키는 방법의 구현은 여전히 어려운 일이다. Independent Cascade (IC) 모델이나 Linear Threshold (LT) 모델을 응용한 정보확산 최대화 또는 최소화 유도가 가능하지만[3], 초기정보를 파악하는 것에는 한계가 있다.

본 논문에서는 온라인 정보의 초기 상태를 분석하기 위한 측정 방법을 제안하고, 자동화된 크롤링 도구를 구현 하였다. 본 논문의 주요 성과는 다음과 같다: (1) 온라인 정보의 초기상태 분석을 위한 분석 기준(토론수준, 긍정도, 민감도, 참여도) 제시한다. (2) 온라인 초기 정보 수집(crawling) 프로그램을 구현하여 온라인 공간에서 실제 정보 수집한다. (3) 서로 다른 정보제공 분야별 특성의 차이점 존재 여부에 대한 분석을 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 기존 연구에 대하여 살펴본다. 3장에서는 초기 정보 파악을 위한 기준을 설정한다. 4장에서는 타겟 시스템을 선정과 데이터 수집 방법에 대하여 설명한다. 수집된 정보의 분석 결과와 정보 분야별 특징의 분석 결과는 5장에서 제시한다. 6장의 결론으로 본 논문을 마무리 한다.

Ⅱ. 관련 연구

인터넷의 발달에 따른 최근의 정보 전달 및 속도에 관한 연구의 시작은 정보 네트워크 분야에서 비롯된 것은 아니었다. 의학 및 보건 분야의 유행성 전염병 감염에 대한 연구들이 정보 전달 연구에 영향을 주게 되었다. 보다 구체적으로, 정보 확산에 관련한 초기 연구는 1990년대 후반 보건 분야에서 전염병(epidemic)의 급격한 확산 속도에 대한 초기 연구[4,5]로부터 시작되었다. 전염병에 전파에 대한 과학적 해석을 위해 SIR (susceptible - infectious - removed) 모델이 제안되었다. 의료/보건 분야에서 제안된 SIR 모델을 기반으로 분기처리모델이 탄생하였고 이를 통해 다양한 형태의 네트워크 정보전달 형태를 분석하게 되었다[6]. 정보확산을 해석하는 기본적인 방법으로 IC 및 LT 모델과 이를 응용한 다양한 해석이 가능하다[3].

초기의 네트워크 정보확산 모델은 다양한 형태로 발전하였다. 다계층(multi-level)을 가지는 모델에서의 정보확산 방법이 연구되기도 하였으며[7], 개별 유저들의 사회적 연결 또는 만남(social contacts)과 상호간의 preferences를 분석하여 미래 특정 시점의 정보확산 정도를 추정하는 모델[8], 정확한 최적값을 정확히 찾지는 못하는 기존의 탐욕(greedy) 알고리즘이 가지고 있는 단점을 보완하기 위한 접근법[9] 등이 연구되고 있다.

급속하게 발달하는 전자상거래에서는 온라인 소셜 네트워크의 구전 효과(word of mouth)를 활용해 마케팅 성과를 극대화 하기 위한 정보 전달 최대화 문제(information diffusion maximization problem)도 연구되었다[10]. 또한 트위터(Twitter)와 같은 초소형 블로그(micro-blog)에서 오피니언 리더를 식별하고 이들을 통한 정보 전달 프레임워크(framework)를 제안하여 온라인 마케팅 성과를 확산시키려는 방법이 제시되기도 하였다[11].

온라인 사회 관계망은 오프라인에서 존재하는 부정적 영향(집단 따돌림, 괴롭힘 등) 뿐만 아니라 새로운 악영향(가짜뉴스 유통, 정보 왜곡/변형, 순위 조작 등)

을 가능하게 만들었다. 특히, 온라인 세상에서의 악영향은 오프라인보다 지속성이 높다는 연구결과가 있다[12]. 이러한 온라인에서의 부정적 영향을 최소화하기 위하여 사용자들이 남긴 글(comments)에 순위를 부여하는 방법[13], 품질이 우수한 사용자 의견을 선별하여 뉴스정보 제공 시스템의 신뢰성을 향상시키는 방법[14] 등이 연구되었다.

그러나 위에서 기술한 연구들은 정보확산, 전자 상거래, 온라인 악영향 감소 등에 관한 연구로 정보의 초기 특성을 분석하여 활용하는 방법에 대한 대안은 제시하지 못하였다. 본 논문에서는 기존의 연구 방향과는 달리 정보 생성 초기 상태를 분석하여 확산/배포와 온라인 부작용의 초기 판단 근거가 되는 정보의 자동화된 분석 방법과 실제 데이터를 활용한 분석을 제시한다.

Ⅲ. 초기 정보 파악을 위한 기준 설정

본 장에서는 인터넷에서 생성되는 정보의 초기 형태를 측정하기 위한 기준을 제시한다. 본 논문에서는 인터넷 정보의 생애 주기(life cycle)를 '생성(birth) - 성장(growth) - 침체(stagnation) - 소멸(death)'의 단계로 정의한다. 정보의 성장 및 확산에 있어 다양한 해석이 가능하지만, 현재까지 '인터넷 정보 생애'를 기반으로 연구가 진행된 것은 드물다. 특히, 인터넷 정보의 생애 주기 관점에서 민감도를 분석하는 방법은 적용된 사례가 거의 없다. 일반적으로 민감도를 측정하는 방법으로는 단일 변수의 변화량의 관찰하는 방법, true positive (TP) rate과 false positive (FP) rate의 상대 비율을 활용하는 ROC/AUC 방법[16,17], Monte Carlo filtering 및 Bayesian uncertainty estimation 등을 활용한 확률·통계적 방법[18] 등이 가능하지만 참과 거짓을 확정할 수 없는 정보 생성 단계에서는 적용이 제한된다. 본 논문에서는 '생성 - 성장' 단계를 분석하는데 초점을 맞추고 민감도 측정 기준을 '초기 반응'시간으로 한정하여 연구를 수행한다. '침체 - 소멸' 단계에 대한 분석은 향후 연구로 남긴다.

3.1 인터넷 정보의 생성-성장 모델

3.1.1 정보 생성(birth)

인터넷 기사(news), 페이스북 또는 개인 블로그 글(posting) 등 인터넷 공간에서는 끊임없이 정보가 생성된다. 이때 특정 시간 t에 생성되는 정보를 i_t로 정의한다.

3.1.2 정보 성장(growth)

i_t가 생성된 이후 다음은 성장의 과정을 거치게 된다. i_t의 성장을 측정하기 위해서는 다양한 관점의 측정 기준이 가능하다. 본 논문에서는 4가지 측정 기준을 제안한다.

긍정도(optimistic level, o_i). i_t가 성장하기 위해서는 긍정적인 반응의 정도가 중요하다. 대부분의 정보제공 시스템에서는 사용자들이 긍정도를 표현할 수 있는 기능을 제시한다. 페이스북의 '엄지(thump-up)', 신문기사의 '좋아요(good)', 블로그 글에 대한 '하트(heart)' 등이 해당된다. 이러한 정보 수집을 통해 i_t의 긍정도를 측정할 수 있다. i_t의 긍정도를 o_i라고 하면, o_i는 긍정 표현의 개수를 의미한다.

토론 수준(discussion level, d_i). i_t에 대하여 i_t에 대하여 다양한 반응이 나타날 경우 원댓글(original reply, 이하 OR) 또는 댓글에 대한 댓글(reply to reply, 이하 RR)을 통해 다양한 토론(discussion)이 발생한다. 토론의 수를 통해 i_t가 성장함을 측정할 수 있다. 본 논문에서는 i_t의 토론 수준을 d_i로 정의한다.

참여도(attendance level, a_i). i_t의 성장은 많은 사람들이 참여할수록 정보 성장의 정도가 달라진다. 참여 수준은 RR의 생성 정도에 따라 측정이 가능하다. 본 논문에서는 참여도를 a_i로 표시하기로 한다.

민감도(sensitivity level, s_i). 얼마나 빠르게 다른 사용자들이 반응하는지에 대한 평가요소로써 민감도를 설정한다. i_t의 민감도를 s_i라 하면, s_i는 i_t의 현재시점($t_{current}$)에서 i_t가 생성된 시점의(t)의 시간 차이의 평균값이다.

3.1.3 정보성장 관측을 통한 초기 특성 파악

초기 인터넷 정보는 정보 생성 직후부터 일정 기간 동안에 발생하는 반응의 정도를 관찰함으로써 정보성장 수준의 측정이 가능하다. 측정 기준은 3.1.2에서 제시한 기준을 활용한다. 그러나 본 논문에서 제시하는 기준은 분석 주체에 따라 추가, 삭제 및 변경이 가능하므로 향후 관련 연구의 확장성을 고려하여 정보성장 관측 모델을 일반화하여 제시한다. 수학적 표현은 수식 (1)와 같다. 수식 (1)에서 u는 측정 기준들의 벡터이다.

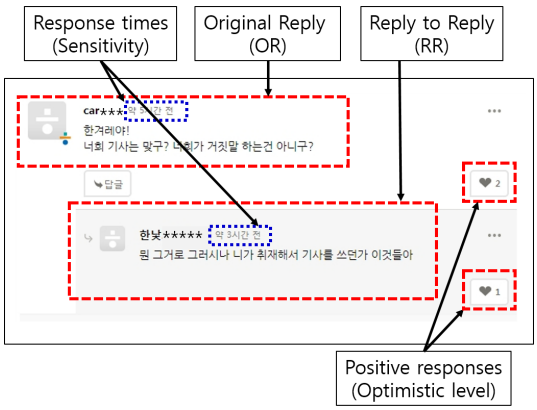

그림 1. OR/RR 구조로부터의 정보 예
Fig. 1. An example of information from OR/RR structure

$$u = (u_1, u_2, \ldots, u_n) \tag{1}$$

본 논문에서 새롭게 제안한 기준으로 한정할 경우 $n = 4$, $u_1 = o_i$, $u_2 = d_i$, $u_3 = a_i$, $u_4 = s_i$, 이다. 각각의 측정 기준은 분석 주체에 따라 서로 다른 가중치를 가질 수 있다. 특정 시스템은 토론수준(d_i)에 높은 가중치를 부여할 수 있고, 다른 시스템에서는 민감도(s_i)에 더 높은 가중치를 부여할 수도 있다. 이러한 상황을 일반화하면 수식 (2)와 같이 표현할 수 있다. 수식 (2)에서 w는 측정 기준의 가중치 벡터이다.

$$w = (w_1, w_2, \ldots, w_n) \tag{2}$$

수식 (2)에서 모든 가중치의 합은 1이며 w 개별 원소의 값은 [0, 1] 사이의 값을 갖는다. 단일 scalar 값을 갖는 정보수준을 산출하기 위해 수식 (1)에서 산출된 값과 가중치를 내적(inner product)한다. 다만 벡터 연산을 위해 w는 전치(transpose) 이후 벡터를 적용한다. 따라서 본 논문에서는 일반화된 정보성장 측정 방식을 수식 (3) 과 같이 제안한다.

$$L_i = u \bullet w^T \tag{3}$$

수식 (3)에서 L_i는 i_t의 정보 성장 수준을 의미한다.

Ⅳ. 타겟 시스템 선정 및 데이터 수집

4.1 타겟 시스템

인터넷 정보는 다양한 분야에서 끊임없이 생성되고 있다. 가장 활발한 분야는 온라인 포털(네이버, 다음), 온라인 뉴스, social network service (SNS, 트위터, 페이스북), 개인 블로그(blog), 온라인 마켓(아마존, 이베이) 등이 대표적이다. 본 논문에서는 온라인 뉴스 제공 서비스 중 하나를 선택하여 초기 정보성장 관측 타겟으로 지정하였다. 뉴스 제공 서비스는 분야별/주제별 정보 구분이 용이하고, 사용자가 의견을 남기거나(OR) 의견에대한 의견(RR)을 남기는 서비스가 제공되므로 정보생성과 성장과정의 추적이 상대적으로 용이하다. 본 논문에 활용된 타겟은 '한겨레 신문'이다(Fig. 2 참조). 플랫폼별(포털, 뉴스, 마켓, 블로그 등)에 대한 정보성장 특성을 분석하는 것은 본 논문의 향후 연구로 남긴다.

그림 2. 본 논문에 활용된 타겟 뉴스 서비스
Fig. 2. The target news service used in this paper

4.2 데이터 수집

타겟 시스템으로부터 정보를 추출하기 위해서는 자동화된 크롤링(crawling) 도구가 필요하다. 본 논문에서는 크롤링 도구를 Python 언어를 활용하여 직접 구현하였다. 주로 활용한 module은 HTML 및 XML에서 데이터를 추출하게 도와주는 Beautifulsoup과 DOM 처리, CSS selector, JSON, Canvas, SVG와 같은 다양한 웹 표준화를 지원하는 PhantomJS 등을 사용하였다.

초기 정보의 성장 특성을 파악하기 위하여 본 논문에서는 2개의 주요 분야(카테고리)를 선정하였다. 첫 번째 분야는 국방 분야이며, 두 번째 분야는 경제 분야이다. 자료 수집을 위해 최초로 분야별 키워드를 선

정한다. 본 논문의 경우 국방 분야는 5개로써 '트럼프(Trumph)', '시진핑(Xi Jinping)', '송영무(Song Young-mu)', '사드(THAAD)', '김관진(Kim, Kwan-jin)'을 선정하였으며, 경제 분야의 5개 키워드는 '무역', '기업', '경제', '금융', '노동'으로 선정하였다.

분야별 키워드를 활용하여 연관 기사를 검색한다. Fig. 3은 국방 분야의 키워드 중 하나인 '트럼프'를 활용하여 관련 정보(연관 기사)를 추출한 결과이다. 키워드를 활용한 관련 정보 각각으로부터 정보성장 특성을 파악하기 위한 값(d_i, o_i, s_i, a_i)을 계산한다. 이후 최종적인 정보 성장 특성을 판단하기 위해 L_i를 계산한다. 분야별 정보성장 관련 데이터 수집 절차는 Table 1에 제시 하였으며, 이러한 과정을 거쳐 수집된 데이터의 통계 현황은 Table 2에 정리 하였다. 초기 정보 성장 관측을 위해 키워드별 초기 10개의 OR(OR에 따른 RR 포함)에 연관된 정보를 모두 수집하였다.

Table 1. Data collection procedure

Step	Work descriptions
1:	Select major keywords in a category
2:	Extract the information regarding d_i, o_i, s_i, and a_i for each keyword
3:	Compute o_i, d_i, a_i, s_i
4:	Computer L_i

키워드	기사순서	기사제목
트럼프	1	시진핑, 트럼프에 283조 경협 선물보따리 풀었다.
트럼프	2	트럼프가 떠난 뒤[김종철 칼럼]
트럼프	3	외교부, 일본 독도새우 항의에 적절치 않다 반박
트럼프	4	트럼프 국회서 자기 골프장 깨알 자랑 미 언론 엽기적
트럼프	5	정부, 평택기지 비용 50% 부담은 거짓이였나
트럼프	6	실익 다 챙기고 떠난 미국... 기업 만나 투자금 확약 받았다
트럼프	7	국정원 수사와 적폐들의 반란 +트럼프 방한 분석
트럼프	8	로봇설 멜라니아는 이럴 때만 광대승천 미소를 짓는다
트럼프	9	반트럼프 시위대 본 미국쪽 의외의 반응... 인기 실감
트럼프	10	따듯한 느낌이 없어서... 즉석에서 만찬사 수정한 트럼프
트럼프	11	[세상 읽기] 근조, 한반도 평화는 죽었다.
트럼프	12	시진핑, 트럼프 위해 자금성 통째로 쓰며 황제의전

그림 3. 키워드 '트럼프'에 대한 검색 결과
Fig. 3. The results of keyword search with 'Trumph'

Table 2. The statistics of the collected data

Category	Defense	Economy
# Related Articles	96	45
# Optimistic Signs (# 'likes')	1,603	506
Ave. Reply to reply (RR)	1,709	603
Ave. length of RRs	228.8	29.5
Ave. hours until 10 replies formed	198.0	21.5

V. 정보 성장 분석

5.1 분야별 분석 접근법

본 논문에서 제시하는 2개 분야별 초기 정보 성장 특성을 분석하기 위하여 Table 2에서 정리된 데이터를 기반으로 분석을 수행하였다. 데이터로부터 일정 수준을 측정하기 위하여 각각의 i_t는 2017년 11월 7일을 기준으로 뉴스 제공 시스템에서 수집된 정보이다(즉, t=2017. 11.7.). o_i는 i_t와 관련 정보(연관 기사)에 생성된 모든 긍정 표시(한겨레 뉴스 제공의 경우 '하트' 모양)의 개수이다. d_i는 연관 기사에 생성된 RR의 개수이며, a_i는 특정 정보(i_t, 본 논문의 경우 키워드)에 대한 연관 기사 중 OR의 개수가 5개 이상 생성된 연관 기사의 비율로 측정하였다. s_i는 특정 정보(i_t)에 대하여 초기 10개의 OR 및 RR이 생성될 때까지 소요된 시간의 평균값이다.

각 측정 기준별 절대점수의 차이, 상대적 표준편차가 존재하므로 종합적으로 비교 판단하기 어렵기 때문에 본 논문에서는 이러한 문제점을 해결하기 위하여 logistic 함수를 적용하였다[18]. 본 논문에서 사용한 logistic 함수는 수식 (4)와 같다. 본문에서 적용한 파라미터 값은 독자들의 직관적 이해를 돕기 위해 측정변수들의 분포 형태를 최소 0점, 최대 100점, 중앙값 50점으로 결과 분포를 재조정(re-formation)하기 위하여 수식 (4)로부터 x_0 = 50, L = 100, k = 0.1로 설정하였다. 모든 측정 기준에 따른 입력값은 [-∞, ∞]의 범위 안에서 가능하며, 결과값은 최소 0, 최대 100, 중앙값 50을 갖게 된다.

$$f(x)= \frac{L}{1+e^{-k(x-x_0)}} \tag{4}$$

여기서 x_0는 중앙값, L은 최대값, k는 곡선의 steepness이다.

본 논문에서 형성되는 입출력 값들의 그래프 형태는 Fig. 4에 제시하였다.

수식 (4)를 활용하여 산출한 o_i, d_i, a_i, s_i 값은 가

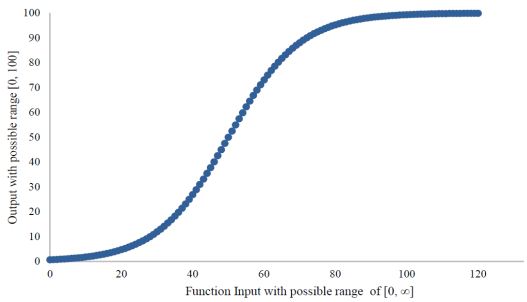

그림 4. 본 논문에 적용된 로지스틱 함수 형태
Fig. 4. The shape of logistic function used in this paper

중치 부여를 위하여 수식 (2)를 적용해야 한다. 가중치의 정확한 산출은 대량의 데이터를 분석한 이후 추가적 leaning 과정 이후에 설정될 수 있을 것이다. 본 논문은 정보생성 단계의 특성을 파악하는 것이 주요 목적이므로 우선 모든 측정 기준의 비중을 동일하게 설정하였다. 정보 출처별 가중치 특성 파악 및 영향력 분석은 본 논문의 향후 연구로 남긴다. 즉 w_1 = w_2 = w_3 = w_4 = 0.25 로 균등하게 설정하였다. 본 절에서 적용한 분야별 분석 접근법을 '국방' 및 '경제' 분야에 적용하여 얻은 특정 사례 값 들을 Table 3, 4에 제시하였다.

5.2 분야별 데이터 분석

국방(defense) 분야의 경우 키워드 '김관진'은 긍정도, 참여도에서 정보의 초기 성장 정도가 큰 것으로 나타났다. 데이터 수집 시점(t = 2017. 11. 7.)에서 키워드 '송영무(국방부장관)'는 토론 수준이 높게, 키워드 '시진핑'은 민감도가 높게 나타났다.

종합점수는 '김관진'이 19.35로 가장 높았다. 결과, 본 논문에서 실제 데이터를 활용한 분석 결과 t를 기준으로 '국방' 분야에서 가장 관심을 가져야할 키워드는 높은 정보성장 수준을 보이는 키워드는 '김관진'이었다.

경제(economy) 분야의 경우 키워드 '무역'이 참여도와 민감도 부분에서 높은 점수를 보여 주었다. 키워드 '기업'은 토론수준에서, '경제'는 긍정도에서 높은 점수를 보였다. 종합점수는 '무역'이 5.69로 가장 높았다. 결과, t 시점의 경제 분야에서 가장 관심을 가져야 할 키워드는 '무역'으로 분석되었다. 국방과 경제 분야의 특성은 다음과 같다. 2개 분야 모두 긍정도와

Table 3. The results of analysis on 'Defense' category

Metric		Optimistic Level (o_i)	Discussion level (d_i)	Attendance Level(a_i)	Sensitivity Level (s_i)	Total Score
Sub key-word	Trumph (i=1)	3.29	5.07	7.59	8.03	5.99
	Xi Jinping (i=2)	2.33	5.61	3.46	12.04	5.86
	Song Yong-mu (i=3)	1.97	10.26	15.84	9.51	9.40
	THAAD (i=4)	3.29	5.07	23.69	8.03	10.02
	Kim, Kwan-jin (i=5)	6.78	1.87	58.18	10.57	19.35
Ave.		3.45	5.22	18.91	11.88	9.87
STD.		1.72	2.38	20.70	5.71	4.96

* 검정 음영은 정보성장 측정 기준 중에서 최대값을 나타냄

Table 4. The results of analysis on 'Economy' category

Metric		Optimistic Level (o_i)	Discussion level (d_i)	Attendance Level(a_i)	Sensitivity Level (s_i)	Total Score
Sub key-word	Trading (i=1)	1.85	2.96	10.25	7.69	5.69
	Enterprise (i=2)	1.97	3.41	5.57	3.48	3.61
	Economy (i=3)	2.52	2.10	3.45	6.95	3.75
	Labor (i=4)	1.66	2.26	3.45	7.21	3.64
	Financing (i=5)	2.79	1.55	1.80	3.65	2.45
Ave.		2.16	2.46	4.90	5.80	3.83
STD.		0.48	0.73	3.27	2.05	1.17

* 검정 음영은 정보성장 측정 기준 중에서 최대값을 나타냄

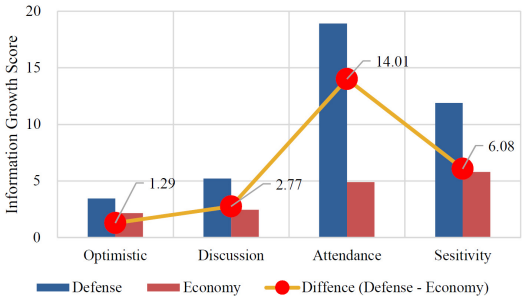

그림 5. 카테고리 사이의 비교 결과 (국방 및 경제)
Fig. 5. The results of comparison between categories (defense and economy)

토론수준은 낮게 나타났으며, 분야 간 차이도 작았다.

참여도와 민감도는 2개 분야가 고루 높게 나타났지만, 국방 분야의 참여도는 5.78배가 높았으며, 민감도는 2.05배 높았다. 국방과 경제 분야의 주요 특성은 Fig. 5에 정리하여 제시하였다.

본 논문에서 분석한 뉴스 제공 서비스의 정보 수용자들의 정보 데이터로 볼 때, 국방 분야가 경제 분야보다 2.58배(초기 정보 성장 지수 기준) 빠르게 성장하는 것으로 나타났다.

5.3 측정기준 상관도 분석

본 논문에서는 측정 기준 사이의 의존도 존재 여부를 파악하기 위하여 추가적인 분석을 실시하였다. 국방 분야의 키워드에 따른 모든 연관기사별 긍정도, 토론수준, 민감도에 대한 연관성을 분석하였다. 각각의 쌍대 비교를 통해 분포도와 결정계수를 확인하였다 (Fig. 6 참조). 분포도의 경우 '긍정도 - 토론수준', '긍정도 - 민감도', '토론수준 - 민감도' 사이에는 특별한 상관관계가 나타나지 않았다. 실제로 각 측정기준 사이의 결정계수를 확인 결과 '긍정도 - 토론수준'은 0.071, '긍정도 - 민감도'는 0.016, '토론수준 - 민감도'는 0.0049로써 각 지표별 상관관계는 없는 것으로 나타났다. 추가적으로 각 지표별 분포에서 차이점이 존재하는 지 여부를 검증하기 위해 t-test를 수행한 결과 p-value는 '긍정도 - 토론수준' 0.00044, '긍정도 - 민감도'는 0.000000009, '토론수준 - 민감도'는 0.000002로 모든 지표들 사이에 유의하게 차이점이 존재하는 것을 확인하였다, 따라서 본 논문에서 제안하는 지표들은 상호 의존 확률이 매우 낮은 가운데, 정보 성장 단계에서 상호 독립적으로 유용하게 사용될 수 있다.

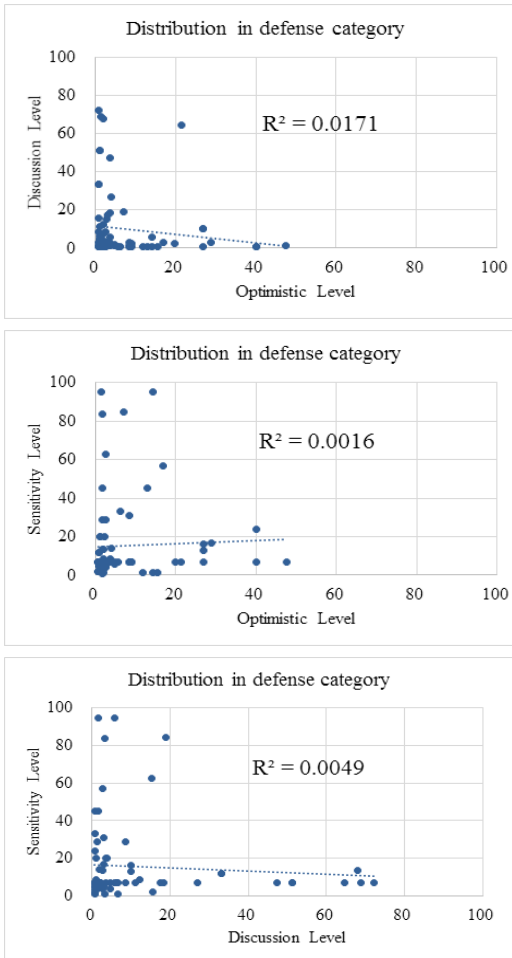

그림 6. 국방 카테고리에서 긍정도 vs. 토론 수준, 긍정도 vs. 민감도, 토론 수준 vs. 민감도 비교
Fig. 6. The comparisons in optimistic vs. discussion, optimistic vs. sensitivity, discussion vs. sensitivity in defense category

Ⅵ. 결 론

본 논문에서는 급격하게 증가하고 있는 인터넷 정보의 효과적인 특성 파악을 위해 '생성 → 성장 → 침체 → 소멸'의 단계로 구분하고 성장 단계의 분석 방법론을 제시하였다. 정보의 성장 단계를 분석하는 것은 인터넷에서 발생하는 가짜뉴스, 온라인 따돌림, 순위 변형/왜곡, 정보 조작 등과 같은 부작용의 초기 단계를 분석하여 정보화 부작용의 초기 단계를 탐지하고 모니터링 하는데 유용하게 활용될 수 있다. 본 논문에서는 4가지 정보 성장 측정 기준을 제시하고 자동화된 크롤링(crawling) 도구를 설계하여 실제 데이터를 수집하였다. 이를 통해 새로운 정보 분석 방법론

을 제시한다.

수집된 실제 데이터를 활용하여 '국방'과 '경제' 분야의 초기 정보 성장 수준을 측정하고 결과를 제시하였다. 데이터 수집 시점을 기준으로 정보 생성이후 '국방'분야의 성장 수준이 2.58배 빠르게 성장함을 확인하였다. 향후 연구로는 생성 정보의 장기적 분석(long-term analysis)과 정보화 부작용의 사례와 연관한 추가 분석이 필요하며, 정보 분야별 가중치 설정, 정보 성장이 정보화 부작용 패턴과 동일한 경우 경고 시스템 설계 등 다양한 분야의 연구가 필요할 것으로 판단된다. 추가적으로 연속 시계열 상에서 자동화된 동적 분석 도구의 설계가 필요하다.

References

[1] Wikipedia, *Cyberbullying*, Retrieved Jan. 30, 2018, from: https://en.wikipedia.org/wiki/Cyberbullying

[2] Wikipedia, *Fake news*, Retrieved Jan. 30, 2018, from: https://en.wikipedia.org/wiki/Fake_news

[3] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo, "The independent cascade and linear threshold models," *Diffusion in Soc. Networks*, Springer, pp. 35-48, 2015.

[4] J. Diamond and C. Renfrew, "Guns, germs, and steel: The fates of human societies," *Nature,* vol. 386, pp. 339-339, 1997.

[5] C. McEvedy, "The bubonic plague," *Scientific American,* vol. 258, pp. 118-123, 1988.

[6] D. Easley and J. Kleinberg, "Networks, crowds, and markets," *Cambridge Univ. Press,* vol. 6, pp. 1-6, 2010.

[7] Y. Zhou, B. Zhang, X. Sun, Q. Zheng, and T. Liu, "Analyzing and modeling dynamics of information diffusion in microblogging social network," *J. Network and Computer Appl.*, vol. 86, pp. 92-102, 2017.

[8] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, "Modeling information diffusion over social networks for temporal dynamic prediction," *IEEE Trans. Knowledge and Data Eng.*, vol. 29, no. 9, 2017.

[9] D.-L. Nguyen, T.-H. Nguyen, T.-H. Do, and M. Yoo, "Probability-based multi-hop diffusion method for influence maximization in social networks," *Wireless Pers. Commun.,* vol. 93, no. 4, pp. 903-916, 2017.

[10] T. Araujo, P. Neijens, and R. Vliegenthart, "Getting the word out on Twitter: the role of influentials, information brokers and strong ties in building word-of-mouth for brands," *Int. J. Advertising,* vol. 36, no. 3, pp. 496-513, 2017.

[11] F. Li and T. C. Du, "Maximizing micro-blog influence in online promotion," *Expert Syst. with Appl.,* vol. 70, pp. 52-66, 2017.

[12] J. Liu, H. Shen, and L. Yu, "Question quality analysis and prediction in community question answering services with coupled mutual reinforcement," *IEEE Trans. Services Comput.,* vol. 10, no. 2, pp. 286-301, 2017.

[13] C.-F. Hsu, E. Khabiri, and J. Caverlee, "Ranking comments on the social web," in *Proc. 2009 Int. Conf. Computational Sci. and Eng.*, vol. 4, 2009.

[14] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist, "Supporting comment moderators in identifying high quality online news comments," in *Proc. 2016 CHI Conf. Human Factors in Computing Syst.*, San Jose, California, USA, 2016.

[15] Wikipedia, *Receiver operating characteristic,* Retrived, Jan. 27, 2018, from https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=822589022

[16] S. W. Song, "Using the receiver operating characteristic (ROC) curve to measure sensitivity and specificity," *Korean J. Family Med.*, vol. 30, no. 11, pp. 841-842, 2009.

[17] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice: a guide to assessing scientific models*, John Wiley & Sons, 2004.

[18] Wikipedia, *Logistic regression*, Retrived Jan. 30, 2018, from: https://en.wikipedia.org/wiki/Logistic_regression

노 기 섭 (Giseop Noh)

1994년 3월 : 공군사관학교 산업공학과 졸업
2009년 8월 : University of Colorado Denver 컴퓨터 과학 석사
2014년 8월 : 서울대학교 컴퓨터공학과 박사
2018년 3월~현재 : 청주대학교 조교수
<관심분야> 인터넷정보분석, 소셜네트워크, 보안

오 하 영 (Hayoung Oh)

2002년 2월 : 덕성여자대학교 전산학 졸업
2006년 2월 : 이화여자대학교 컴퓨터 공학 석사
2013년 2월 : 서울대학교 컴퓨터공학과 박사
2013년 9월~2016년 8월 : 숭실대학교 조교수
2016년 9월~현재 : 아주대학교 조교수
<관심분야> 소셜 정보망 및 유무선 네트워크