

2016. 1. 15.

빅데이터를 활용한 소셜네트워크 분석 SW 관련 인식, 화제성, 연관어 추이 비교 연구

이동현 (dlee@spri.kr)

(1) 연구의 목적

- 소셜네트워크(SNS) 및 뉴스기사 빅데이터를 활용하여 SW에 대한 인식 및 이슈내용이 어떻게 변화되고 있는지 살펴보고자 함
 - SW에 대해 ‘누가’ ‘언제’ ‘어디서’ ‘어떤’ 이야기를 ‘왜’ 하고 있는지 파악하고 이를 계량화
 - SW에 대한 인식을 지수화함으로써 SW중심사회의 현주소를 점검하는 진단도구로 사용

(2) 분석방법

- 국내의 대표적인 SNS 서비스중 트위터, 네이버 블로그, 대형 커뮤니티 게시판 그리고 인터넷 뉴스를 수집해놓은 빅데이터¹⁾에서 ‘소프트웨어’와 연관된 텍스트를 추출
 - 트위터: 연간 약 18억건
 - 네이버 블로그: 연간 약 1.5억건
 - 국내 140여개 대형 커뮤니티 (보배드림, 닥취고취업, 82cook, mlb파크, 뽀뿌, 미즈넷, 네이트판 등)
 - 국내 110여개 언론사 뉴스 (종합, 경제, IT, 방송/통신, 인터넷, 지역, 매거진 등 네이버 인터넷 뉴스 전체)
- 본 연구에서는 분석을 위하여 3가지 기간으로 나누어서 ‘소프트웨어’에 대한 화제성, 인식도 및 연관어 추이를 비교
 - 각 기간 별 포함된 일수를 동일하게 343일²⁾로 설정하여 분석함

1) 다음소프트 제공

2) 최초 연구 시점과 SW중심사회 선포식 사이의 일 수

<표 1> 추이분석을 위한 기간설정

번호	기간	비고
1	2012/04/14 ~ 2013/03/22	미래부 출범이전
2	2013/08/15 ~ 2014/07/22	SW중심사회 선포식 이전
3	2014/07/23 ~ 2015/06/30	SW중심사회 선포식 이후

(3) 화제성

- 화제성: SW를 언급한 소셜네트워크 및 뉴스에 대한 빈도 수
- SW중심사회 선포식 이후 전체 ‘소프트웨어’ 언급 횟수는 증가하여, 긍정적으로 분석됨
 - 다만, 언론에서는 SW중심사회 이전에 비하면 SW에 대한 화제성이 떨어지고, 미래부 출범 이전에 비해도 소폭 상승에 그쳐 소셜네트워크 위주로 화제성이 올라간 것으로 보임

<표 2> 화제성 지수

(단위: 건)

채널	미래부 출범 이전	SOS ³⁾ 선포식 이전	SOS 선포식 이후
소셜 네트워크	191,735	184,239	206,437
뉴스	23,912	25,862	23,988
전체	215,647	210,101	230,425

(4) 인식도

- 인식도: SW와 함께 사용되는 감성어 중 긍정어/부정어의 비중
 - SW가 포함된 문장을 형태소 분석, 구문 분석, 개체명 인식, 품사 태깅, 복합명사처리 등의 언어학적 지식등을 이용하여 단어들을 추출

3) SOS(Software-Oriented Society): 소프트웨어 중심사회

- 형태소 사전을 이용하여 추출된 자연어를 긍정어/중립어/부정어 로 구분

□ SW 대한 부정적 인식은 점차 줄어드는 추세 (17.6% → 15.8% → 14.0%)

- 긍정적 인식은 ‘비율’ 기준으로는 증가한다고 보기 힘드나, ‘건수’ 기준으로는 증가하는 추세 (66,846단어 → 78,399단어 → 91,923단어)

<표 3> 인식도 지수

인식도	미래부 출범 이전		SOS 선포식 이전		SOS 선포식 이후	
	건수	비율	건수	비율	건수	비율
긍정	66,846	43.4%	78,399	40.5%	91,923	41.4%
중립	58,544	38.0%	82,717	42.7%	97,435	43.9%
부정	27,079	17.6%	30,486	15.8%	31,172	14.0%

- ‘소프트웨어’ 에 대한 언론기사의 부정적 표현의 횟수 및 비중이 점차 줄어드는 추세 (6,894단어 (16.1%) → 5,111단어 (11.3%) → 3,763단어 (11.2%))

<표 4> 인식도 지수 - “뉴스”

인식도	미래부 출범 이전		SOS 선포식 이전		SOS 선포식 이후	
	건수	비율	건수	비율	건수	비율
긍정	18024	42.0%	20,400	44.9%	14,477	43.2%
중립	17,609	41.0%	19,429	42.8%	15,004	44.8%
부정	6,894	16.1%	5,111	11.3%	3,763	11.2%

□ 긍정어: 긍정적인 표현에 주로 쓰이는 단어

- 소프트웨어와 관련된 긍정어의 예: 탁월한, 뛰어난, 세계최초, 혁신적, 안전한, 빠른, 최적, 효율적, 최고, 좋은, 개선된, 성공, 창의적, 즐기다, 높은 수준, 향상되다, 성장하다, 안정적, 정교하다, 만족하다, 도움 등

□ 부정어: 부정적인 표현에 주로 쓰이는 단어

- 소프트웨어와 관련된 부정어의 예: 부작용, 불법, 오류, 복잡하다, 어렵다, 불만, 한계있다, 우려, 부족하다, 위험, 위협, 실패하다, 고통, 취약한, 심각하다, 피해, 암울하다, 느리다, 장애, 정보유출, 오작동 등

- 중립어: 상황에 따라 긍정적/부정적 양 쪽에 모두 사용가능한 단어
 - 예시: 다양한, 새로운, 다르다, 필요한, 전망, 지속적, 필요하다, 대표적, 공개하다 등
- 기타(분석에서 제외된 단어들): 감각 및 색상을 나타내는 단어들
 - 예시: 짜다, 블루, 취하다, 화이트 등

(5) 연관어

- 연관어: SW와 함께 사용되는 단어들 중 인물, 단체, 장소, 상품, 속성, 브랜드, 취미/여가, 심리 등을 나타내는 명사
 - ‘소프트웨어’에 대한 연관어 중 가장 눈에 띄는 것은 ‘교육’으로써, 순위가 급격히 상승 중 (40위 → 29위 → 16위)
 - ‘하드웨어’, ‘개발’, ‘프로그램’, ‘컴퓨터’를 전체 연관어 순위가 “뉴스”만 놓고 분석했을 경우와 비교해서 상대적으로 높음. 이들 연관어들에 대한 ‘소프트웨어’ 관련 의견피력이 SNS상에서 보다 활발히 이루어지고 있는 것으로 분석됨

<표 5> SW 연관어 순위

연관어 순위	미래부 출범 이전		SOS 선포식 이전		SOS 선포식 이후	
	단어	횟수	단어	횟수	단어	횟수
1	기술	32410	정보	34848	정보	37872
2	기업	30242	기술	32312	기술	36354
3	개발	29563	개발	30604	개발	31563
4	하드웨어	29118	시스템	28082	기업	28525
5	시스템	28234	하드웨어	28045	컴퓨터	27856
6	정보	27257	기업	27619	시스템	27350

7	제품	26891	컴퓨터	27483	하드웨어	25368
8	기능	25460	서비스	25413	서비스	23004
9	서비스	24772	제품	23850	프로그램	22287
10	컴퓨터	24177	기능	22918	제품	21865
11	프로그램	22777	프로그램	22183	기능	21725
12	시장	22410	시장	22114	시장	19915
13	업체	20708	데이터	19658	데이터	19157
14	사용자	18695	업체	19648	인터넷	18810
15	회사	18464	사용자	17467	업체	18759
16	데이터	17139	산업	16859	교육	17722
17	산업	16616	회사	16756	산업	17655
18	스마트폰	16236	사업	16239	사용자	17579
19	삼성	15788	스마트폰	15887	삼성	16966
20	환경	15626	지원	15685	회사	16598
21	운영	15462	운영	15183	지원	16450
22	애플	15457	보안	15063	사업	15556
23	지원	14830	인터넷	14886	운영	15069
24	모바일	14292	모바일	14777	미국	15002
25	사업	14168	환경	14130	스마트폰	14708
26	인터넷	13835	기기	14023	환경	14325
27	미국	13764	삼성	13558	보안	14139
28	pc	13642	게임	13296	모바일	13967
29	보안	13139	교육	12635	개발자	13833
30	게임	12980	미국	12325	기기	13297

- SNS에 비해서 교육의 연관어 순위가 낮은 편이지만, ‘소프트웨어 교육’에 관련된 언론기사 역시 점차 많아지는 추세 (연관어 순위: 103위 → 42위 → 35위)
- ‘기업’, ‘시장’, ‘산업’, ‘서비스’, ‘업체’, ‘스마트폰’, ‘모바일’의 연관어 순위가 뉴스만 놓고 분석했을 때 상대적으로 높은 것으로 보아, 이들 연관어에 대한 ‘소프트웨어’ 관련 언론기사가 많이 작성되는 것으로 추정

<표 5> SW 연관어 순위

연관어 순위	미래부 출범 이전		SOS 선포식 이전		SOS 선포식 이후	
	단어	횟수	단어	횟수	단어	횟수
1	기업	9984	기업	11116	기업	9538
2	기술	9524	기술	10169	기술	9266
3	시장	8851	서비스	9532	정보	8487
4	서비스	8031	시장	9320	서비스	7736
5	개발	7425	정보	8937	개발	7466
6	정보	7380	개발	8655	시장	6966
7	업체	7228	업체	7490	업체	6099
8	제품	6945	제품	6894	사업	5984
9	시스템	6332	시스템	6634	시스템	5411
10	모바일	5976	모바일	5903	인터넷	5391
11	스마트폰	5432	사업	5852	산업	5051
12	기능	5112	스마트폰	5630	제품	4886
13	산업	4678	산업	5254	계획	4014
14	사업	4591	기능	5237	환경	3870
15	솔루션	4449	기기	4699	스마트폰	3743
16	회사	4306	솔루션	4500	기능	3718

17	하드웨어	4034	데이터	4428	지원	3642
18	게임	3934	계획	4403	글로벌	3598
19	사용자	3901	보안	4386	데이터	3543
20	데이터	3868	인터넷	4201	운영	3349
21	보안	3850	하드웨어	4185	솔루션	3347
22	환경	3734	사용자	4145	모바일	3316
23	운영	3673	지원	4144	하드웨어	3217
24	기기	3660	글로벌	4141	회사	3201
25	업계	3440	회사	4029	기기	3184
26	애플리케이션	3248	운영	3759	사용자	3164
27	애플	3241	삼성	3624	콘텐츠	3085
28	네트워크	3237	환경	3582	플랫폼	3054
29	글로벌	3231	업계	3519	미국	3044
30	삼성	3198	플랫폼	3340	보안	2897

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.