공학석사 학위논문

KeyplayerRank를 이용한 소셜 미디어상의 주제별 키플레이어 탐지기법 연구

A Study on Detection of Subject-based Key Player on Social Media Using KeyplayerRank

2018년 8월

서울대학교 대학원

건설환경공학부

김 민 선

KeyplayerRank를 이용한 소셜 미디어상의 주제별 키플레이어 탐지기법 연구

A Study on Detection of Subject-based Key Player on Social Media Using KeyplayerRank

지도교수 김 용 일

이 논문을 공학석사 학위논문으로 제출함

2018년 5월

서울대학교 대학원

건설환경공학부

김 민 선

김민선의 공학석사 학위논문을 인준함

2018년 5월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

국문초록

소셜 미디어 이용자가 매년 꾸준히 증가하면서 소셜 미디어를 이용하여 자료나 정보를 습득하고 소셜 미디어상에서 다른 사용자와 커뮤니케이션 하는 사람들이 늘어가고 있다. 자료나 정보를 습득할 때 소셜 미디어에서 가장 영향력 있는 사용자의 말을 신뢰하고 따르는 경우가 빈번하게 일어나고 있다. 따라서 본 연구에서는 특정 주제에 관심 있는 사람들 간의 관계에 가중치로 작용하는 주제 유사도와 사용자의 영향력과 파급력을 고려한 영향력 지수를 적용한 KeyplayerRank를 이용하여 소셜 미디어상의 키플레이어를 찾고자 한다. 본 연구에서는 트위터를 대상으로 주제별 키플레이어를 탐지하였으며, 주제를 추출하기 위해 토픽모델링 방법 중 LDA를 이용하였고, 주제와 관련 있는 텍스트만을 분류하기 위해 SVM을 이용하였다.

본 연구에서는 두 가지 지표를 동시에 고려하여 소셜 미디어상에서 키플레이어를 찾고자 하는 목적에 따라 사용자가 선택할 수 있는 새로운 방법을 제안하였다. 제안한 방법론은 트위터만이 아닌 다른 소셜 미디어에도 적용 가능하며, 기업의 인플루언서 마케팅이나 여론형성을 통한 정책 결정에도 활용될 수 있다.

향후 본 연구는 시공간적 요소를 포함하여 실시간으로 각 주제가 이슈가 되었던 때의 실시간 키플레이어를 찾는 연구로 발전시킬 수 있다.

주요어 : KeyplayerRank, PageRank, LDA, SVM, 키플레이어, 소셜 미디어
학 번 : 2016-21243

목 차

표 목 차

그 림 목 차

1. 서론

1.1 연구 배경 및 목적

인터넷과 스마트 기기를 활용하여 자신을 표현하고 소통하는 소셜 미디어(social media)의 이용률은 매년 꾸준히 증가하고 있다. 김윤화(2015)에 따르면 2015년 전체 응답자 9,873명 중에서 4,250명이 소셜 미디어를 이용하고 있다고 응답하였고, 16.8%에 그치는 2011년 소셜 미디어 이용률과 비교하면 2015년 소셜 미디어 이용률은 43.1%에 달했다.

또한, 2013년 한국인터넷진흥원에서 모바일인터넷 이용실태 조사를 한 결과 모바일인터넷의 이용 목적에서 자료 및 정보 습득이 95.9%로 가장 높았고, 커뮤니케이션이 94.6%로 그다음으로 높았다(그림 1-1).

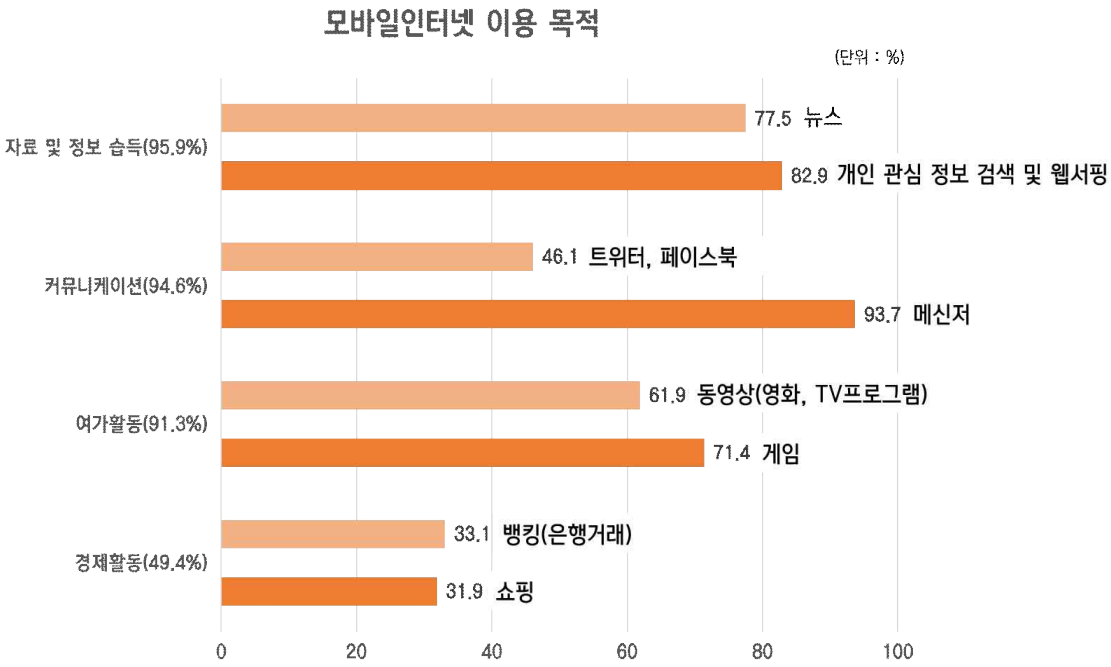

그림 1-1. 모바일인터넷 이용 목적(한국인터넷 진흥원, 2013)

실제로 미국의 여성들을 대상으로 소셜 미디어 트렌드를 분석하는 기업인 BlogHer에서 조사한 결과 소셜 미디어상에서 영향력이 있는 사람의 정보를 주로 습득하고, 상품을 구매할 때도 영향력이 있는 사람의 말을 신뢰하는 것으로 나타났다. 그리고 최근 기업들도 상품을 개발하고 홍보하는데 다중 채널 네트워크(MCN, Multi Channel Network)[1]에서 영향력 있는 사람을 활용하고 있다. 이처럼 다양한 분야에서 소셜 미디어의 사용자 관계 구조 및 네트워크 내에서 영향력을 행사하는 사람을 찾는 것은 점차 중요해지고, 그에 관한 연구도 활발히 진행되고 있다.

연구 및 자문 기업인 Altimeter의 보고서의 '영향력의 3 기둥(Pillars of Influence)'으로 영향력이 있는 사람의 '영향력'을 정의할 수 있다. 첫 번째 기둥은 도달성(reach)으로, 소셜 그래프에서 형성된 관계가 소셜 그래프와 커뮤니티 전체에 정보를 얼마나 멀리 이동시킬 수 있는지를 정의한다. 도달성을 측정하는 척도로 인기도, 근접성, 호의가 있다. 두 번째 기둥은 연관성(relevance)으로, 관심 그래프와 그 커뮤니티를 이어주는 주제적 관련성을 의미한다. 주제적 관련성을 알기 위해선 신뢰도, 친밀감, 권위성으로 판단 할 수 있다. 마지막 기둥은 공명성(resonance)으로, 도달성과 연관성의 "점수"의 기초가 된다. 공명성은 빈도, 기간, 네트워크의 연계 정도로 측정된다. 예를 들어, 트위터에서의 영향력을 영향력 기둥으로 설명한다면, 믿고 따르는 팔로워(follower) 수가 많고 특정 주제와 관련성을 가지며, 그 주제에서 리트윗(retweet)이 많이 되는 것을 의미한다.

다른 사람에게 이러한 영향력을 끼치는 자를 '영향력자' 혹은 '키플레이어(key player)'라고 부른다. 본 연구는 소셜 미디어 데이터 중 트위터를 이용하여 특정 주제에서의 키플레이어를 찾아내고 그 결과를 순위로 나타내고자 한다.

---

1) 다중 채널 네트워크(MCN)는 2014년 전 유튜브 직원 제드 시몬스에 의해 명명되었으며, 1인 미디어의 등장으로 관련 시장이 주목을 받으면서 그들을 하나로 모아 체계적인 관리를 통해 시너지를 극대화하는 것을 의미함.

본 연구는 특정 주제에서의 키플레이어를 찾기 위해 주제 유사도(Topical Similarity)와 영향력 지수(Influence Index)를 고려하였다. 주제 유사도는 사용자들 간의 주제적 유사성을 고려하는 것으로 특정 주제에 관심 있는 사람들 간의 관계에 가중치로 이용한다. 영향력 지수는 사용자의 파급력, 영향력을 측정하는 지표로 이 지표의 값이 커질수록 이 사용자의 영향력은 커진다. 본 연구에서 두 가지 지표를 고려한 이유는 아래와 같다.

먼저 주제 유사도를 고려한 이유는 다음과 같다. Kleinberg(1998)에 의하면 단순히 in-degree[2]가 높은 사람은 다양한 사람들이 팔로우하기 때문에 사용자가 찾는 특정 주제가 아닌 지나치게 다양한 주제를 언급한다고 한다. 여기서 in-degree가 높다는 것은 많은 수의 사용자들이 한 사용자를 팔로우하는 것을 의미한다. 이러한 점은 특정 주제에 관심 있는 사용자가 그 주제의 키플레이어를 찾을 때 관련되지 않은 주제들을 걸러내야 하는 어려움을 겪을 수 있다. 따라서 주제와 관련 있는 사용자 중에서 키플레이어를 찾는 것이 필요하다. McPherson *et al.*(2001)에 의하면 사람들은 자신과 비슷한 사람들과 교류하려고 하므로 사람들이 얻는 정보, 활동 및 상호작용에 강력한 영향을 주어 사람들의 사회적 세계를 제한하기도 한다. 이것은 유사한 사람들 간의 관계가 유사하지 않은 사람들 간의 관계보다 더 밀접하므로 생기는 현상이다. 따라서 트위터에서 키플레이어를 찾을 때 주제 유사도를 이용하는 것은 같은 관심사를 가진 사용자들 사이의 관계에서 어떤 사용자가 같은 주제 내의 다른 사용자에게 더 영향력을 끼칠지를 알 수 있다. Weng *et al.*(2010)에 의하면 현재 트위터와 다른 많은 소셜 미디어 어플들에서는 한 사용자의 영향력을 측정할 때 그 사용자가 가지고 있는 팔로워 수로 결정한다고 한다. 또한, 트위터 데이터를 분석한 결과 사용자의 72.4%가 자신의 팔로워의 80% 이상을 팔로잉하고, 80.5%에 해당하는 사용자가 80% 이상의 맞팔률[3]을 보인다고 밝혔다. 따라서 이러한 결과는 팔로잉의 관계

2) 방향 그래프에서 한 정점으로 들어오는 연결선의 수, 내차 혹은 입력 차수로도 불림(정보통신기술용어해설).
3) 사용자 간에 서로 팔로우를 하는 것.

가 단순히 자신을 팔로우한 사람을 예의상 팔로우하는 것으로 볼 수 있다고 분석하였다. 그러므로 특정 주제에 관심 있는 사용자 간의 관계를 분석하기 위해선 친구 간의 예의상 맞팔이 아닌 주제 유사도를 이용해야 한다.

그리고 키플레이어를 탐지하는 데 있어 영향력 지수를 고려한 이유는 다음과 같다. Leavitt *et al.*(2009)에 의하면 키플레이어는 잠재적 행동을 통해 다른 사람의 행동을 유도하는 사람을 의미한다. 다른 사람의 행동을 유도한다는 것은 그만큼 다른 사용자에게 영향력을 끼치고 있는 것을 말한다. Krigsman(2010)은 진정한 소셜 미디어 키플레이어는 팔로워 수만 많은 사람이 아니라 다양한 측면에서 많은 사용자와 광범위하게 상호작용하는 사람이라고 정의하였다. 그리고 키플레이어의 판단이나 충고를 다른 사용자들이 신뢰할 때, 이들은 진정한 '상호 연결적 리더(nexus leader)'라고 보았다.

Malcom Gladwell은 그의 저서 <티핑포인트(The Tipping Point)>에서 항상 새로운 지식과 정보를 습득하고 공유하기를 좋아하는 사람들을 '메이븐(Maven)'이라 칭하고, 어떤 제품을 대유행시키려면 메이븐을 통한 새로운 입소문 구전효과를 주목해야 한다고 강조했다. 여기서 분류한 키플레이어들은 특정 주제에 전문가적 식견을 지닌 사람들이 중심이다(Malcom, 2000). 이처럼 본 연구를 통해 트위터의 특정 주제에서 가장 영향력 있는 키플레이어를 찾는 것을 통해 소셜 미디어에서 사용자들이 관심을 가지는 주제에 대한 정보를 탐색하는 데에 소모되는 시간을 줄일 수 있다.

또한 KeyplayerRank(이하 KpRank) 알고리즘을 단순히 트위터에만 적용하지 않고 다른 소셜 미디어에도 적용해 볼 수 있다는 점에서 의의가 있다. 또한, 본 연구에서는 두 지표를 동시에 고려하기 때문에 영향력 지수와 주제 유사도를 따로 구하여 더하는 것과는 다른 방식의 키플레이어 탐지 방법을 제안한다. 영향력 지수와 주제 유사도를 단순 결합하여 키플레이어를 도출할 경우, 주제 자체가 가지는 영향력에 매우 의존적인 결과가 도출될 수 있다. 즉, 키플레이어를 탐지하고자 하는 주제

가 많은 사람으로부터 관심을 받는 주제가 아닌 경우에 대해서는, 해당 주제의 키플레이어가 다른 주제의 키플레이어에 비해 상대적으로 낮은 영향력 지수를 가지게 된다. 이 경우 주제 유사도가 높음에도 불구하고 낮은 영향력 지수로 인해 원하는 주제에 대한 키플레이어를 도출할 수 없게 된다. 반대로 영향력 지수가 아주 높은 사용자여도 사용자가 관심 있는 주제의 키플레이어가 아닐 수 있다. 따라서 본 연구에서는 두 지표를 한쪽에 편향되지 않게 동시에 반영하여 키플레이어를 탐지하는 것을 목적으로 한다. 본 연구를 통해 주제별 상대적인 키플레이어를 찾을 수 있다. 그리고 주제별로 가장 영향력 있는 사용자를 통해 주제에 대한 확실한 정보를 얻을 수 있으며, 불필요한 정보는 거를 수 있다는 장점이 있다.

향후 본 연구를 발전시켜 키플레이어를 중심으로 실시간으로 형성되는 여론을 파악할 수 있고, 더불어 인플루언서 관계 마케팅에도 이용될 수 있다.

1.2 연구 동향

사회적 영향력에 관한 연구가 점차 중요해지게 된 것은 Katz and Lazarsfeld(1955)에 의한 '2단계 흐름 모형(two-step flow model)'에서 시작되었다. 2단계 흐름 모형은 수많은 개인이 미디어와 대중을 연결짓는 소수의 '의견 지도자(opinion leader)'들을 따름으로 미디어에 직접 설득되지 않는다는 것을 강조하고 있다(그림 1-2). 의견 지도자들은 정보 제공자의 입장으로 사람들에게 기존에 알려지지 않은 정보를 제공하는 것을 즐기며, 이미 사람들이 알고 있는 정보에 대해서도 다른 사람들과 함께 논의하려는 경향을 보인다(이원태, 2011).

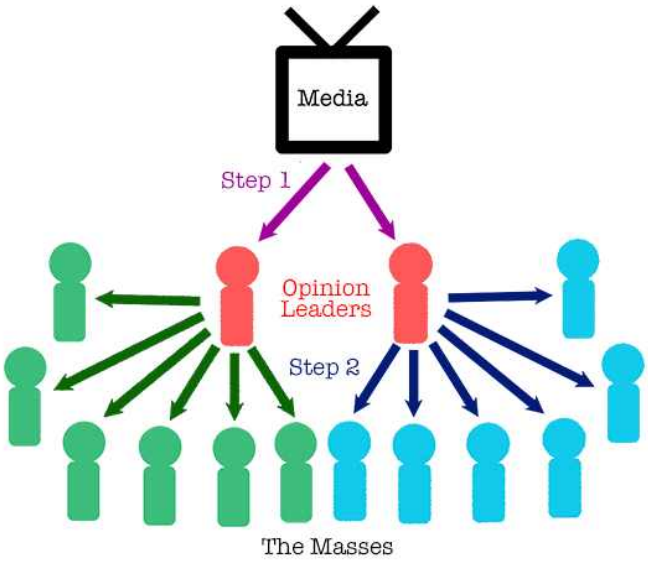

그림 1-2. 2단계 흐름 모형
(Katz and Lazarsfeld, 1955)

소셜 미디어가 발달하면서 사회적 영향력을 소셜 미디어에서 규명하려는 시도는 최근 다양한 방법으로 이루어지고 있다. 특히 소셜 미디어 중에서도 트위터에서 영향력 있는 자를 찾으려는 연구가 가장 활발하다.

먼저, Java *et al.*(2007)은 트위터를 대상으로 소셜 네트워크의 위상 및 지리적 특성을 연구하여 마이크로 블로깅 현상을 관찰하였다. 이를 위해 HITS(Hyperlink-Induced Topic Search) 알고리즘을 사용하여 상위 10개의 허브와 그에 해당하는 권위자 계정을 파악하였다. 또한, 네트워크 분석을 통해 사람들이 마이크로 블로그를 하는 목적은 일상적인 활동에 관해 이야기하고 정보를 찾고 공유하는 것임을 알아내었다.

Cha *et al.*(2010)은 트위터 사용자의 팔로워 수, 리트윗 수, 멘션 수 등 다양한 기준을 세워 5천 200만 명의 트위터 사용자와 연관된 2천만 개의 팔로우 링크와 1,700만 개의 트윗을 비교하며, 트위터에서 누가 영향력 있는 자인가를 탐색하였다. 이 중 적극적인 트위터는 극히 일부이며 트위터 사용자들은 불균형한 영향력을 가지고 있어 트위터 내의 영향력과 명성은 항상 일치하지 않기 때문에, 단순히 팔로워 수가 많다고 하여 인플루엔셜[4]이라고 할 수는 없다는 결론을 내렸다.

비슷한 연구로 Kwak *et al.*(2010)은 '유효 수용자(effective reader)'라는 개념을 도입하여 정보의 확산에 있어 정보를 수용하는 순서가 중요하다는 전제하에 인플루엔셜을 규명하였다. 여기서 '유효 수용자'는 같은 문맥의 정보를 이전에 접해보지 않은 새로운 정보 수용자를 의미한다. 팔로워 수와 '유효 수용자'의 수를 비교하면 트위터 사용자의 80%가 자신의 팔로워의 20%만을 '유효 수용자'로 가지고 있어, 단순히 팔로워 수만 많다고 인플루엔셜이라고 할 수 없다는 사실을 증명하였다.

따라서 소셜 미디어에서 영향력을 측정하는 기준은 한 가지가 될 수 없으며, 소셜 미디어에서 영향력을 측정하기 위한 여러 기준의 측정 도구들이 생겨나고 있다(표 1-1).

4) 영향력자 또는 키플레이어.

표 1-1. 소셜 미디어 영향력 측정 도구 사례

측정 도구	출시 연도	영향력 평가 기준
트윗얌/ 넷다이버	2009	가입자에 한해 해당 트위터의 트윗 수, 팔로워 수, 팔로워가 가진 팔로워 수
Twitalyzer	2009	해시태그, 리트윗 수, 트위터 계정 언급 수, 하이퍼링크 유무 등
Social Media Analytics/SAS	2010	팔로워 수, 트위터 메시지 빈도, 리트윗 수
연합뉴스/ 연합뉴스 미디어랩	2011	해당 사용자의 트윗을 수집해 많이 언급된 키워드를 한눈에 보여줌

트윗얌은 트위터의 미디어 가치를 가격으로 나타내주는 서비스이다. 한국어, 영어, 일본어를 지원하며 달러, 유로, 엔, 위안, 원화로 가격이 표시되어 전 세계인들의 트위터 계정 가치를 가격으로 알 수 있게 제공하고 있다. 계정 가격은 트위터 사용자의 트윗 수, 팔로워 수, 팔로워가 가진 팔로워 수를 고려하여 책정된다. Twitalyzer는 해시태그, 리트윗 수, 트위터 계정 언급 수, 하이퍼링크 유무 등을 분석하여 사용자의 영향력을 측정하고, SAS는 팔로워 수, 트위터 메시지 빈도, 리트윗 수를 고려하여 그래프로 트위터 사용자의 영향력을 나타낸다. 연합뉴스 미디어랩은 해당 사용자가 트윗한 내용을 수집하여 가장 많이 언급된 키워드를 워드클라우드(word cloud) 형식으로 나타내준다.

최근에는 구글이 사용하고 있는 페이지랭크(PageRank)를 소셜 미디어에 적용하여 영향력자를 찾으려는 연구들이 진행되고 있다(표 1-2). 페이지랭크는 여러 링크로 서로 연결된 인터넷 웹페이지들의 네트워크에서 어떤 페이지가 가장 중요한지 알아내는 알고리즘이다.

표 1-2. 소셜 미디어에 페이지랭크를 적용한 연구

저자	연구주제	연구에 이용된 지표	
		주제 유사도	영향력 지수
Tunkelang (2009)	A Twitter Analog to PageRank		●
Weng et al. (2010)	TwitterRank: Finding Topic-sensitive Influential Twitterers	●	
Sung et al. (2013)	The Influence in Twitter: Are They Really Influenced?	●	
박호성 외 (2010)	소셜 네트워트에서의 인플루엔셜 랭킹		●

Tunkelang(2009)는 트위터에서 영향력을 판별하는데 키플레이어가 리트윗하는 내용을 포함한 모든 트윗을 팔로워들이 읽을 확률로 계산하는 TunkRank를 적용하였다. 그러나 TunkRank는 트위터의 내용을 고려하지 않고 리트윗에 중점을 둔 알고리즘이다.

Weng *et al.*(2010)은 TwitterRank를 제안하였다. 기존의 소셜 네트워크 분석 방법에서 키플레이어를 찾을 때 in-degree를 이용하여 단순하게 팔로워 수로 영향력을 측정하는 것은 영향력 개념을 정확하게 나타내지 않으며 페이지랭크는 전체적인 네트워크의 연결 구조를 고려하여 in-degree를 개선하였지만, 트위터 사용자 간에 영향력을 주는 사용자 간의 관심도를 고려하지 못했다고 지적하였다. 따라서 트위터 사용자 간의 연결 구조와 주제 유사도를 고려하여 in-degree와 페이지랭크의 단점을 보완하고자 TwitterRank를 적용하였다. TunkRank와는 반대로 TwitterRank 알고리즘은 주제 유사도에 중점을 둔 Rank 방식을 개발하였다.

Sung *et al.*(2013)도 트위터 사용자 간의 주제적 유사성을 고려한 InterRank를 제시하였다. TwitterRank와 다른 점은 코사인 유사도를

트위터 사용자 간의 관계에 가중치로 곱하였다. InterRank 알고리즘도 TwitterRank와 같이 주제 유사성만을 고려한 방법을 제안하였다.

박호성 외(2010)도 구글의 페이지랭크 알고리즘을 트위터에 적용하여 트위터의 인플루엔셜을 찾는 방법을 소개한 바 있다. 트위터 사용자를 웹페이지로 생각하고 팔로우 관계를 링크로 생각하여 적용하면, 팔로우 수가 많을수록 페이지랭크가 높은 경향이 있어 팔로워만을 가지고 인플루엔셜을 측정하였을 때와 구성은 비슷하지만 인플루엔셜 의미의 차이에 의해 결과가 달라진다고 하였다.

만약 주제 유사성만을 고려한 키플레이어를 찾고 싶다면 TwitterRank와 InterRank를 이용하여 찾을 수 있다. 반대로 유명인사, 정치인 등과 같이 특정한 주제에 제한하지 않고 영향력이 있는 사용자를 찾고 싶다면 TunkRank와 박호성 외(2010)의 알고리즘을 이용할 수 있다.

본 연구에서는 선행연구들과 같이 페이지랭크 알고리즘을 적용하되 주제 유사도와 영향력 지수를 동시에 반영하여 트위터 사용자들 간의 주제적 유사성과 영향력 정도를 동시에 고려하는 알고리즘을 제안한다. 만약 두 가지를 따로 구하여 더한다면 다음과 같은 모순이 생긴다. 예를 들어 영향력 있는 사용자가 한 주제에 관해 이야기하지 않고 여러 주제를 조금씩 언급만 한다면 그 주제에 전문적인 지식이 없다고 보고 키플레이어라고 할 수 없다. 그런데도 선행연구들은 이 사용자를 키플레이어라고 부른다. 이러한 키플레이어는 영향력에만 편향된 키플레이어라고 할 수 있다.

하지만 본 연구에서는 두 가지를 따로 구하여 더하는 방법이 아닌 두 가지를 동시에 고려하여 키플레이어를 찾기 때문에 한쪽에 편향되지 않는 키플레이어를 찾을 수 있다.

1.3 연구 범위 및 방법

본 연구에서는 소셜 미디어 데이터 중 트위터를 이용하여 키플레이어를 탐지하였다. 트위터는 사용자들이 140자 이내 짧은 단문 메시지(트윗)를 작성하여 자신의 트위터 웹에 업데이트하여 개인의 의견이나 생각을 공유하고 소통하는 무료 소셜 네트워킹 및 마이크로 블로깅 서비스이다. 트위터의 뜻은 새가 지저귀는 소리(tweet)를 뜻하는 영어 낱말을 이용하였다. 트위터의 뜻과 같이 트위터는 사용자들이 일상의 작은 얘기들을 수시로 올릴 수 있는 소셜 미디어 공간을 의미한다.

본 연구에서 키플레이어를 찾는데 트위터 데이터를 이용한 이유에는 세 가지가 있다. 첫 번째로 트위터는 오픈 소스(open source)[5]이며 공공데이터로 다양한 사람들의 네트워크를 분석할 수 있다. 두 번째로 트위터는 다른 소셜 미디어와 달리 특정한 제약 없이 자유롭게 친구 관계를 맺을 수 있다. 일부 다른 소셜 미디어들은 사용자가 친구 관계를 요청한 다른 사용자에게 친구 링크를 허용해야 친구 관계가 맺어지는 제약이 있다. 그러나 트위터의 경우 단순하게 팔로우 관계만으로 친구 관계를 형성할 수 있다. 마지막으로 트위터는 다양한 콘텐츠들을 연합한 채널이기 때문에 주제를 분석하는 데 있어 폭넓은 소스를 제공해 준다(Simply Measured, 2014).

본 연구에서 KpRank를 이용하여 주제별 트위터 키플레이어를 탐지하는 방법은 그림 1-3과 같다.

첫 번째 단계에서는 2015년 7월 22일부터 2016년 7월 22일까지 약 1년간의 조선일보 기사를 스크래핑하여 주제를 추출하기 위해 수집한다. 주제를 추출하는데 뉴스 기사를 이용하는 이유는 트위터의 특성상 작성할 수 있는 텍스트가 140자 이내로 한정되어 있어 주제를 추출하기에 부적합하기 때문이다(Livne *et al.*, 2010). 그리고 트위터는 언론 뉴스에 즉각적으로 반응하며 대중의 의견을 신속하게 파악하는 소셜 미디어이

5) 일정한 이용 조건을 지키면 누구나 개량하고, 재배포할 수 있도록 무상으로 공개되는 소스 코드(source code)(한국정보통신기술협회 IT 용어사전).

기 때문에, 뉴스 기사에서 추출한 주제를 트위터에 적용할 수 있다(진설아 외, 2013). 따라서 트위터와 같은 기간의 뉴스 기사를 수집하여 참고 데이터로 사용한다. 추출된 기사에서 기사의 내용을 담고 있는 'TEXT' 필드만을 추출하여 전처리 과정을 거친다. 전처리 과정은 기사 텍스트를 분석하기 위해 문장 단위로 분리한 후 단어별로 tokenize 시키고, 그 후 텍스트에서 한자, 영문자, 숫자와 같은 불용어를 제거하여 형태소 분석을 하는 과정을 말한다. 형태소 분석을 마치면 품사가 태그된 텍스트에서 명사만을 추출한다. 추출된 명사는 Support Vector Machine(SVM)을 통해 주제와 관련 있는 내용과 관련 없는 내용으로 분류된다. SVM을 통해 스크래핑한 기사 내용이 주제와 관련성이 있는지 아닌지를 자동으로 판별하고, 다음 단계인 토픽모델링(topic modeling) 결과의 성능을 높이기 위해서 텍스트를 분류할 수 있다. SVM 결과 기사 내용이 주제와 관련 있다고 검증된 텍스트만을 가지고 토픽모델링을 하였다. 토픽모델링은 원본 텍스트의 단어를 분석하여 해당 텍스트를 통과하는 주제를 추정하고, 이 주제들이 서로 연결되는 방법 및 시간이 지남에 따라 변경되는 방식을 발견하는 통계적 방법이다(Blei, 2012). 본 연구에서는 토픽모델링 방법 중 가장 대표적인 LDA(Latent Dirichlet Allocation)를 이용하여 주제를 추출하였는데, LDA는 다른 토픽모델링 방법들과 비교했을 때 성능이 가장 뛰어났기 때문이다(강애띠, 2016). LDA로 주제별 키워드가 추출되면 해당 키워드를 기반으로 주제를 판별할 수 있다. 추출된 키워드는 임시로 저장하여 트위터에서 사용자를 추출하는 데 이용된다. REST(Representational State Transfer) API(Application Programming Interface)로 수집된 2015년 7월 22일부터 2016년 7월 22일까지의 트위터에서 추출된 키워드를 언급하는 사용자와 텍스트 그리고 사용자의 속성 정보를 가져온다. 가져온 트위터 데이터에서 텍스트 분석을 하기 위해 'text' 필드만을 가져와 전처리 과정을 진행한다. 전처리 과정은 앞서 조선일보 뉴스 기사 텍스트를 전처리했던 것과 같이 진행한다. 명사만 뽑힌 트위터를 가지고 SVM으로 한 번 더 텍스트 분류를 한다. 이때 사용하

는 SVM 모델은 이전에 뉴스 기사를 분류했던 모델에 트위터 데이터를 추가한 모델을 이용한다. 트위터에서도 텍스트 분류를 하는 이유는 조선일보 뉴스 기사의 텍스트를 분류한 것과 같이 주제와 연관성 있는 텍스트만을 추출하기 위함이다. 트위터에서 주제별 키워드에 해당하는 텍스트들을 뽑을 때 주제와 관련 없는 텍스트가 뽑히는 경우가 많다. 예를 들어, '행사'라는 단어가 포함되는 텍스트를 뽑아온다고 하면 '여행사진을 모으고 있어요.'와 같은 텍스트가 추출되기도 한다. 이러한 오류를 줄이기 위하여 텍스트 분류를 하였다.

두 번째 단계에서는 주제별 사용자 분류를 하고 가중네트워크를 생성한다. 키워드를 언급한 사용자 중에는 리트윗하는 사용자들이 대부분이었다. 리트윗을 한 사람의 속성 정보 중 RT_count는 리트윗된 사람의 RT_count가 나오기 때문에 이러한 사용자는 따로 분류 과정을 수행하였다. 자세한 과정은 2.3.1절에서 설명하고 있다. 사용자를 분류한 후 사용자들의 'text', 'followers_count', 'RT_count', 'favorite_count' 필드를 이용하여 주제 유사도와 영향력 지수를 구한다.

마지막 단계에서는 앞서 구한 주제 유사도와 영향력 지수를 이용하여 최종 KpRank를 구한다. KpRank는 기존의 페이지랭크를 변형한 식으로, 특정 주제에서 다른 사용자와의 주제 유사도가 유사하고 영향력 지수가 높은 사람을 키플레이어로 탐지해주는 개념이다. 페이지랭크와 같이 키플레이어와 연결될수록 영향력 있는 사람이 된다. KpRank의 최종 결과물은 순위로 나타나는데 이때 1순위인 사람을 키플레이어라 할 수 있다.

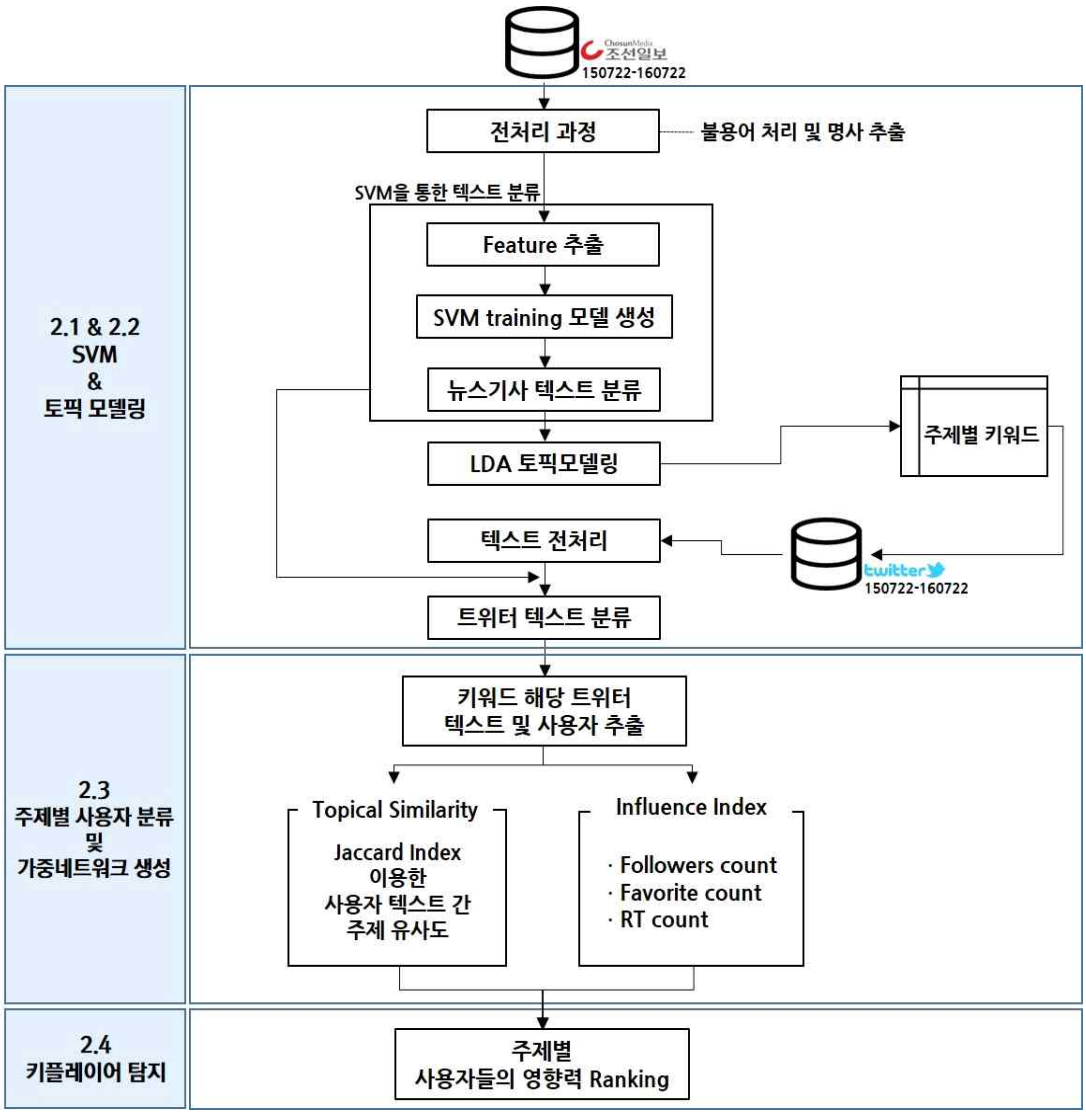
조선일보
150722-160722
2.1 & 2.2
SVM
&
토픽 모델링
전처리 과정
불용어 처리 및 명사 추출
SVM을 통한 텍스트 분류
Feature 추출
SVM training 모델 생성
뉴스기사 텍스트 분류
주제별 키워드
LDA 토픽모델링
텍스트 전처리
twitter
150722-160722
트위터 텍스트 분류
2.3
주제별 사용자 분류
및
가중네트워크 생성
키워드 해당 트위터
텍스트 및 사용자 추출
Topical Similarity
Jaccard Index
이용한
사용자 텍스트 간
주제 유사도
Influence Index
· Followers count
· Favorite count
· RT count
2.4
키플레이어 탐지
주제별
사용자들의 영향력 Ranking

그림 1-3. 연구의 흐름도

2. KeyplayerRank 분석기법

2.1 SVM을 통한 텍스트 분류

본 연구에서는 주제와 관련이 있는 내용만을 선별하기 위하여 text를 분류하였다. text를 분류하기 위해 문서를 자동으로 분류해주는 알고리즘인 Support Vector Machine(SVM)을 이용하였다.

SVM은 부정예제로부터 긍정예제를 분류하는 결정면(decision surface)을 찾아내는 분류모형으로, 이원 패턴인식 문제를 해결하기 위해 1995년 Vladimir Vapnik에 의해 제안된 머신러닝 알고리즘이다(Cortes & Vapnik, 1995). SVM은 구조적 리스크 최소화 원칙(Structural Risk Minimization)을 바탕으로 하여 일반화 오류를 줄여주기 때문에 이미지 분류나 텍스트 분류에 있어 딥러닝 못지않은 뛰어난 성능을 보여주고 있다.

Joachims(1998)에 의하면 텍스트를 분류하는데 SVM이 적합한 이유에 대해 다음과 같이 설명하였다. 첫 번째로, 텍스트는 무수히 많은 특징(feature)으로 구성되어 있다. 따라서 차원이 무한으로 커지기 때문에 데이터에 과적합 된 모델이 생성될 수 있다. 하지만 SVM은 이러한 큰 차원을 조절할 수 있는 변수가 존재하므로 텍스트 분류에 적합하다. 두 번째로, 특징 선택에 있어 무관한 특징을 구분해 낼 수 있다. SVM은 무관한 특징을 구별해내기 위해 여러 특징을 학습하여 모든 정보를 고려하여 문서를 분류할 수 있다. 마지막으로, 대부분의 텍스트는 선형적으로 분류할 수 있다. 또한, SVM은 기본적으로 선형적인 분류 방법을 찾기 때문에 텍스트 분류에 적합하다.

SVM에는 긍정예제와 부정예제를 선형적으로 분리할 때 적절한 최대 마진(margin) 분류기를 구축하는 선형 SVM과 선형적으로 분리가 어려운 경우 커널(kernel) 기법을 통해 비선형 결정함수를 만들어 최적의 초평면을 구하는 비선형 SVM으로 구분된다(정영미 외, 2000). 여기서 커널 기법이란, 주어진 데이터를 고차원 특징 공간으로 사상하는 것을 말한다.

선형 SVM은 데이터를 선형으로 분리하는 최적의 선형결정 경계를 찾는데, 이때 서로 다른 데이터들을 최대의 마진으로 분류하는 결정경계 또는 분리 초평면을 찾는 것이다(그림 2-1).

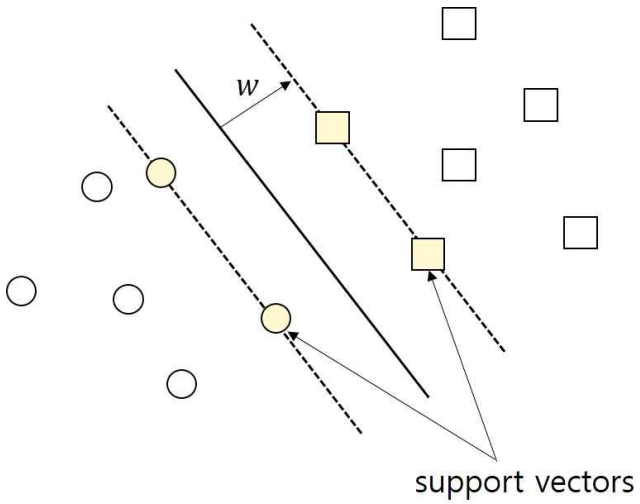

그림 2-1. 선형 SVM

그림 2-1에서 실선은 긍정예제와 부정예제를 분리하는 결정면이고 실선과 평행인 점선들을 마진이라고 한다. 여기서 마진이란 오류를 발생시키지 않으면서 결정면을 움직일 수 있는 공간을 의미한다. 마진이 최대화됐을 때 점선에 있는 데이터들은 결정면으로부터 $\frac{1}{\| w \|}$의 거리에 위치하게 되는데 이것을 support vector라고 부른다(정영미 외, 2000). SVM은 선형적으로 분리할 수 있는 문제를 기반으로 하는데, 선형 분리가 가능하다는 것은 트레이닝 데이터를 긍정예제와 부정예제 두 가지로

나눌 수 있는 것이며 이것을 나누는 결정면이 존재한다는 것을 의미한다 (Dumais *et al.*, 1998). 이 결정면을 수식으로 나타내면 식 2-1과 같다.

$$\vec{w} \cdot \vec{x} - b = 0 \tag{2-1}$$

여기서 $\vec{w}$는 가중치 벡터를 나타내고, $\vec{x}$는 분류하고자 하는 문서의 벡터, b는 기준치로 $\vec{w}$와 b는 트레이닝 데이터로부터 학습된다. 만약 어떠한 문서의 집합을 $D=\{(x_i, y_i)\}$라고 한다면, x_i가 범주에 속하면 y_i는 +1의 값을 가지고, 범주에 속하지 않는다면 -1의 값을 가진다(식 2-2).

$$\begin{aligned} \vec{w} \cdot \vec{x_i} - b \geq +1 \\ \vec{w} \cdot \vec{x_i} - b \leq -1 \end{aligned} \tag{2-2}$$

위의 두 식을 결정함수로 표현하면 다음 식 2-3과 같다.

$$f(x) = sign((w \cdot x_i) + b) \tag{2-3}$$

하지만 모든 데이터가 선형으로 분리되지 않을 경우도 있는데, 이때는 커널(kernel)을 사용하여 선형으로 분리되지 않는 데이터들을 특정한 방식으로 변수 공간을 확장한다(그림 2-2). 즉 주어진 데이터를 고차원 특징 공간으로 사상해주는 것이다.

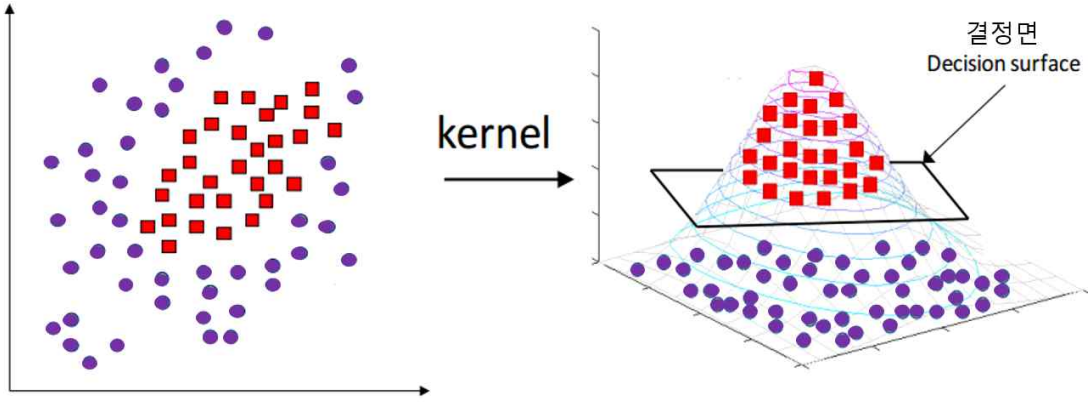

그림 2-2. Kernel로 변수 공간을 확장시킨 모습

커널에는 Polynomial 커널, Sigmoid 커널, 가우시안 RBF 커널 등이 있다. Scholkopf *et al.*(1997)과 Romero *et al.*(2015)에 따르면 이 중 분류 성능이 좋은 커널 함수는 가우시안 RBF 커널이다. RBF 커널은 gamma라는 매개변수를 조정하여 결정면을 최적화한다. gamma는 하나의 데이터 샘플이 영향력을 행사하는 거리를 결정하며, 가우시안 함수의 표준편차와 관련되어 있다. 즉 gamma가 클수록 한 데이터 포인터들이 영향력을 행사하는 거리가 멀어지는 반면 gamma가 낮을수록 그 거리는 짧아진다. 또한, C값도 같이 조정해줘야 하는데, C는 cost(C)를 나타내며 얼마나 많은 데이터 샘플이 다른 클래스에 놓이는 것을 허용하는지를 결정한다. C가 낮을수록 많이 허용하고, 작을수록 적게 허용한다. 따라서 좋은 성능을 얻기 위해 gamma값과 C값을 적절하게 조정하여 최적의 매개변수 값을 지정하는 것이 중요하다.

본 연구에서는 주제와 관련 있는 데이터만을 정확하게 추출하기 위하여 가우시안 RBF 커널을 이용하였다. 그리고 적정한 gamma값과 C값을 결정하기 위하여 교차 타당성 평가를 통해 최적의 매개변수 값을 설정하였다.

2.2 토픽모델링(Topic Modeling)

2.2.1 토픽모델링 관련 연구

토픽모델링(Topic Modeling)이란 데이터 마이닝 기법의 하나로 비구조화된 텍스트 자료의 뭉치에서 의미 있는 주제들을 추출해주는 확률 모델 알고리즘이다. 이전에는 Latent semantic analysis(LSA) 기법이 사용되었으나, Blei *et al.*(2003)이 Latent Dirichlet Allocation(LDA) 알고리즘을 발표한 이후로는 주로 LDA 기법 혹은 LDA의 변용 기법들이 사용되고 있다(남춘호, 2016). LSA는 선형 기법인 특이치 분해 Singular Value Decomposition(SVD)을 통해 행렬의 분해 및 차원 축소를 수행하여 대용량 데이터 집합의 분류를 효율적으로 수행하고, 새로운 의미 공간을 생성하여 특징들의 중의적 의미를 분석하여 분류 대상의 잠재적 의미를 알아낼 수 있다. 하지만, LSA의 차원 축소는 전체 데이터의 표현 정도만을 고려할 뿐 분류하고자 하는 범주를 고려하지 않으며, 서로 다른 범주 간의 차별성을 고려하지 않기 때문에 축소된 차원상에서 다른 범주 데이터 간의 모호한 경계로 인해 안정된 분류 성능을 나타내지 못한다는 한계가 있다(김정호 외, 2010). 반면, LDA는 주어진 문서에 대해 각 문서에 어떤 주제들이 존재하는지에 대한 확률 모형으로서 미리 알고 있는 주제별 단어 수 분포를 바탕으로 주어진 문서에서 발견된 단어 수 분포를 분석함으로써 해당 문서가 어떤 주제들을 함께 다루고 있을지를 예측할 수 있다. 따라서 LDA를 이용한 토픽모델링 관련 연구들이 활발히 진행되고 있으며, LDA를 이용한 경우 분류 성공률이 높고 LSA보다는 해석이 용이하다는 장점이 있다.

최근 경영학, 인문사회학 등 다양한 학문 분야에서도 이슈문서 도출 및 응용을 위하여 토픽모델링이 적용되고 있다. 가장 많이 적용되고 있는 분야 중 하나는 논문의 초록을 분석하여 주제들을 추출하고 저자를 식별하거나, 시간의 흐름에 따른 주제 분포의 변화를 통해 해당 학문 분야의 연구 동향을 파악하는 것이다.

박자현 외(2013)는 국내 문헌정보학의 연구 동향을 분석하기 위해 문헌정보학 주요 학술지인 정보관리학회지, 한국문헌정보학회지, 한국비블리아학회지의 1970년도부터 2012년까지 발표 논문 초록을 수집하여 LDA 기반의 토픽모델링을 수행하였다. 토픽모델링을 수행함으로써 문헌 정보학자들이 관심을 갖는 주요 연구주제를 파악하고, 계량정보학의 내용분석에 새로운 방법론을 제시하였다.

김형석 외(2016)는 딥러닝 분야 4,800여 편의 논문을 대상으로 LDA 토픽모델링과 네트워크 분석을 수행하여 2015년까지의 연구 동향을 파악하였다. 토픽모델링을 통해 토픽분포의 변화량을 기반으로 연도별, 논문 유형별 연구 동향을 확인하고, 토픽 간의 유사도를 기반으로 네트워크 구성 및 군집화를 하였다.

LDA 알고리즘뿐만 아니라 다양한 토픽모델링의 성능을 비교 분석한 선행연구들도 있는데, Sriurai(2011)은 토픽모델링을 이용한 분류모형의 성능을 비교하였다. 이때 비교한 분류모형은 나이브베이즈(Naïve Bayes), 의사결정나무(Decision tree), SVM이다. 텍스트 분류에 있어 특징으로 대표되는 bag of words(BOW)와 LDA를 통해 비슷한 특징끼리 묶어준 것을 가지고 분류 모형을 비교하였을 때, 토픽모델과 SVM을 이용하였을 때가 BOW 모델을 이용했을 때보다 약 11.1% 향상된 정확도를 보였다.

김하진 외(2014)는 저자명 모호성 해소를 위해 토픽모델링 기법을 사용하여 저자명을 식별하였는데, 기존 LDA 방식이 용어 특징만을 고려하였다면 본 연구에서는 제3의 메타데이터 특징을 활용하여 Author-Conference Topic Model(ACT) 모델과 Dirichlet-multinomial Regression(DMR) 토픽모델링을 대상으로 저자명 식별 성능을 평가 및 비교하였다. 연구결과 전반적으로 DMR 토픽모델링의 저자명 식별 성능이 ACT 모델보다 우수하였다.

김규하 외(2015)는 5개년도 학회지 영문초록을 이용하여 LDA 방법과 Correlated topic model(CTM) 방법의 Variational expectation maximization(VEM) 추론방법을 통해 엔트로피 값을 비교분석 하였다.

연구결과 LDA에서는 일부 논문에서 사용되는 특정한 단어들이 많이 추출된 반면, CTM에서는 여러 논문에서 공동으로 나타나는 단어들이 많이 추출되었다.

또한, 소셜 미디어 데이터가 급격하게 증가하면서 대용량의 텍스트 데이터인 소셜 미디어 데이터의 주제 분석을 위해 국내외 연구에서 토픽모델링을 많이 활용하고 있다.

Schwartz *et al.*(2013)은 트위터 데이터의 텍스트에 반영된 웰빙에 대한 토픽을 구분해내기 위해 LDA 모델을 적용하였으며, LDA를 사용하여 트위터에서 파생된 동시 출현 단어 세트, 표준 인구 통계 및 사회경제적 지표(나이, 성별, 민족, 소득 및 교육) 이상으로 삶의 만족도를 예측하는 정확도가 향상되었다.

Ghosh *et al.*(2013)은 지오태그 된 트위터 데이터의 텍스트에 잠재된 비만과 관련된 토픽을 찾아내기 위해 LDA 모델을 사용하였고, 공간정보분석을 통하여 지역별 주제에 해당하는 공간패턴을 보여주었다.

국내에서도 소셜 미디어를 이용한 토픽모델링 분석 연구가 수행되었는데, 차윤정 외(2015)는 스마트폰에 대한 트위터 데이터를 활용하여 토픽모델링분석을 통해 사용자 의견에 대한 주요 주제를 선정하고 이를 분석함으로써 마케팅 전략의 방향을 도출하고자 하였다.

진설아 외(2013)는 트위터 데이터를 분석하여 토픽의 변화 시점 및 패턴을 파악하고, 특정 상품명에 관한 키워드를 추출한 뒤 동시출현단어분석(co-word analysis)을 이용하여 토픽과 관련된 키워드를 노드와 엣지로 표현하였다. 연구결과 트위터를 통해 부정적인 언론 뉴스를 접했을 때 트위터에서 즉각적인 반응이 일어나고 빠르게 정보를 확산시키는 역할을 한다는 것을 확인하였다.

하지만 소셜 미디어를 이용한 선행연구들은 공통적으로 소셜 미디어 데이터의 특성상 제한적인 글자 수를 가지기 때문에 LDA 입력값으로 글의 길이가 충분하지 않다는 한계를 보여주었다.

2.2.2 LDA를 이용한 토픽모델링

본 연구에서는 토픽모델링 기법 중 가장 성능이 좋은 LDA를 이용하여 토픽모델링을 수행하였다.

LDA는 그림 2-3과 같이 설명할 수 있다. 여기서 M은 말뭉치 안의 문서를 의미하고, N_d는 문서 안에 있는 단어를 의미한다. α는 문서당 주제 혼합 분포를 의미하고, θ_{dk}는 해당 문서에서 각 주제의 가중치를 의미한다. z_{dn}은 할당된 토픽을 의미하고, w_{dn}은 문서 내에 분포된 단어들을 의미하며 β_{ki}는 토픽당 단어의 분포를 의미한다. 따라서 다음과 같은 과정을 거치는데, 먼저 어떤 문서에 대해 파라미터인 주제 벡터 θ_{dk}가 있고 앞에서부터 단어를 하나씩 채울 때마다 θ_{dk}로부터 하나의 주제를 선택하고, 다시 그 주제로부터 단어를 선택하는 방식으로 문서 생성과정을 모델링 한다.

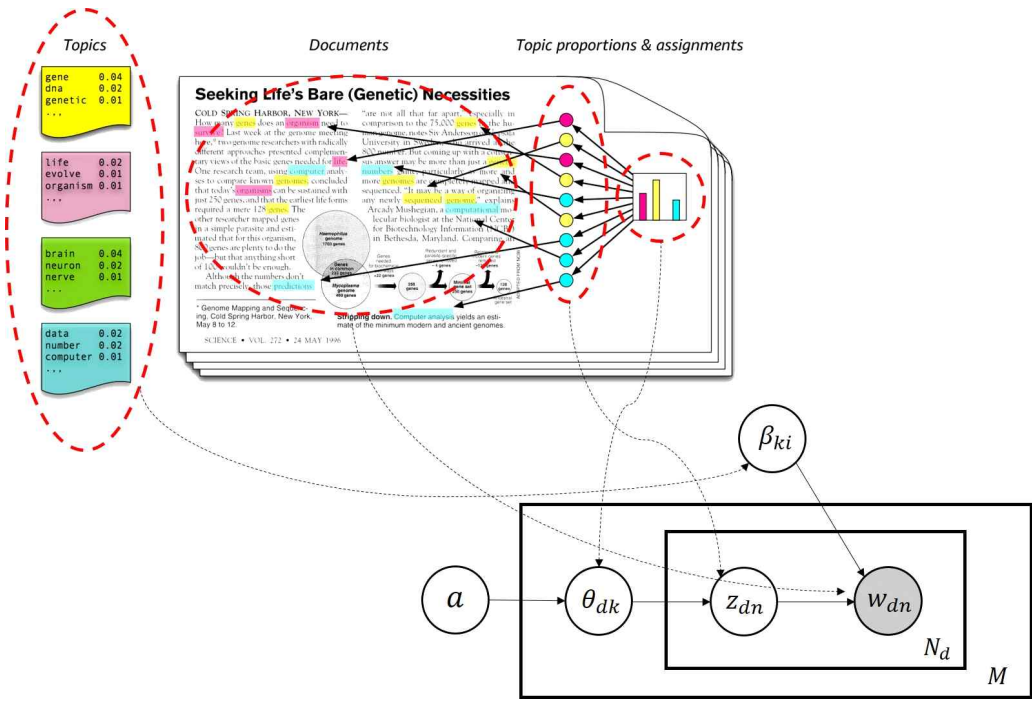

그림 2-3. LDA 알고리즘

본 연구에서는 조선일보 기사에 LDA를 적용하여 주제를 추출하고자 한다. 조선일보 기사는 장문으로 구성되어 있고 트위터와 달리 맞춤법이 잘 갖춰져 있으며, 띄어쓰기가 제대로 되어있어 LDA로 주제를 추출하기가 트위터보다 용이하다. 따라서 본 연구에서는 트위터의 기간에 맞춰 조선일보 기사를 참고 데이터로 사용하여 LDA를 통해 주제를 추출하고자 한다. LDA를 하기 위해 먼저 뉴스 기사에서 형태소 분석을 한 텍스트에 SVM을 적용하여 1차로 관련 있다고 분류된 텍스트를 추출한다. 그다음 명사만을 가지고 말뭉치를 형성하고, 형성된 말뭉치에 속하는 단어에 숫자를 부여하여 각 문장당 말뭉치 단위를 생성한 후 이 데이터를 LDA 모델의 입력 데이터로 사용한다.

LDA 결과 주제별 키워드와 키워드가 해당 주제에 분포할 확률이 함께 나온다. 주제별 키워드를 가지고 트위터에서 해당 키워드를 언급한 사용자를 분류하고 가중네트워크를 생성한다.

2.3 사용자 간 네트워크 생성

2.3.1 주제별 트위터 사용자 분류

이전 단계에서 LDA를 통해 추출된 주제의 키워드를 포함하는 트위터들을 모두 추출하고, SVM을 통해 주제와 관련되는 트위터만을 가져온 후 사용자 분류를 통해 사용자 간의 네트워크를 생성하고자 한다.

추출된 트위터는 'text', 'user_id', 'user_screen_name', 'followers_count', 'retweet_count', 'favorite_count' 필드로 이루어져 있다. 먼저 트위터의 내용을 포함하는 'text' 필드를 보면 트위터의 텍스트들은 단순히 한글로만 이루어져 있지 않다. 예를 들면, 표 2-1과같이 한 트위터 사용자가 작성한 트위터 텍스트가 있다. 여기서 'RT'란 리트윗으로, 아래의 텍스트를 작성한 사용자가 다른 사용자의 글을 자신의 팔로워들에게 전달했다는 뜻이다. 'RT' 옆에 '@사용자 ID'가 있는데 이것은 본 텍스트를 처음으로 작성한 사용자 아이디로 즉, 이 사용자가 'iamkep**'라는 사용자의 글을 리트윗했다고 할 수 있다. 리트윗은 트윗을 손쉽게 재전달하여 빠른 정보 전파를 가능하게 한다(곽해운 외, 2011). 그리고 '#'은 해시태그(hashtag)로 이 글의 키워드를 작성하고 싶을 때 주로 이용된다. 해시태그는 트위터를 찾을 때 글을 쉽게 찾아주는 검색어 기능도 가지고 있다. 또한, 맨 뒤의 URL은 해당 트위터의 출처를 말한다.

표 2-1. 트위터 텍스트 예시

RT @iamkep**: 어제 내린 비로 조금 추워진 출근길... 오늘은 이세돌 대 알파고의 바둑 첫 대결이 있는 날이네요. 인간 대 인간지능... 과연 누가 이길까요? #이세돌 #알파고 #바둑 https://t.co/woChryFwta

이처럼 트위터에는 한 사용자가 직접 작성한 텍스트도 많지만, 다른 사용자가 한 사용자의 글을 리트윗하여 전달하는 경우가 훨씬 많이 일어나고 있다.

트위터 API에서 트위터 정보를 수집해 올 때 리트윗을 한 사용자의 속성 정보는 리트윗된 사람의 속성 정보가 수집되어온다. 위의 표 2-1을 작성한 사용자가 'A'라고 하면 저 텍스트와 관련된 'followers_count', 'retweet_count', 'favorite_count'는 모두 'iamkep**'의 속성 정보가 수집되어온다.

본 연구에서 사용된 트위터에는 리트윗한 트위터가 리트윗하지 않은 트위터보다 2배가량 많았다. 따라서 리트윗을 한 사용자와 리트윗을 하지 않은 사용자를 분류할 필요가 있다. 그러므로 이 단계에서는 다음 그림 2-4와 같은 과정을 수행한다. 이 단계에서 분류해야 할 사용자는 ① 원 글 작성자, ② 리트윗한 사용자, ③ DB 내에 있는 리트윗된 사용자, ④ DB 내에 없는 리트윗된 사용자로 나뉜다.

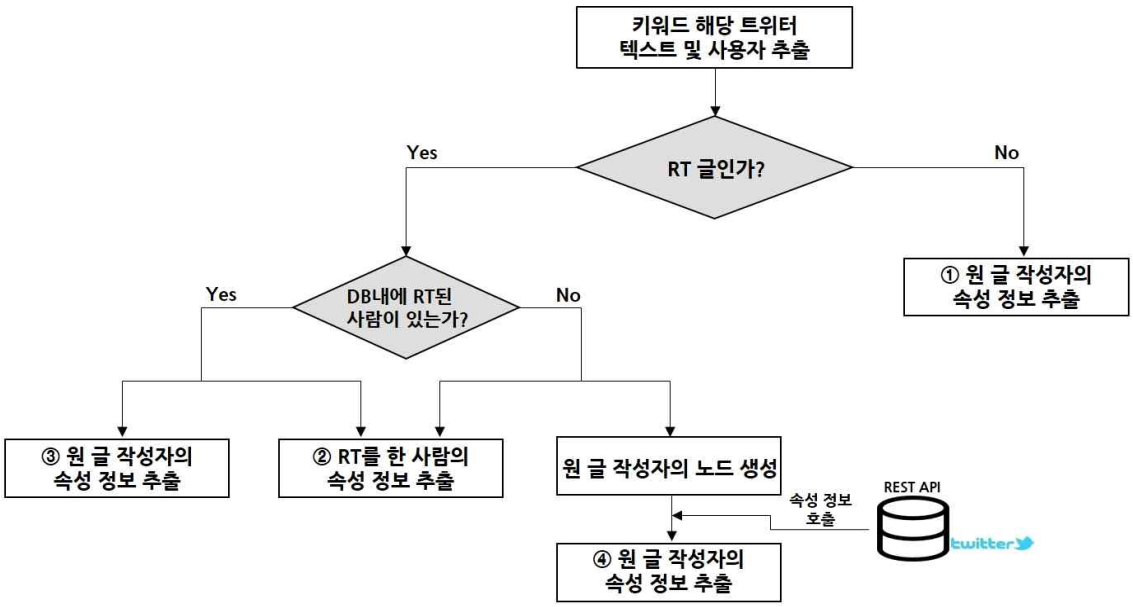

그림 2-4. 트위터 사용자 분류 과정

먼저 작성된 트윗이 리트윗인지 아닌지를 판별한다. 만약 리트윗이 아니라면 작성자 아이디(①)의 속성 정보를 추출한다. 반대로 리트윗된 글이라면 리트윗된 사용자가 본 연구의 트위터 DB에 존재하는지를 판별한다. 트위터 DB에 있는 사용자(②)라면 해당 사용자의 속성 정보를 추출한다. 하지만 DB 내에 리트윗된 사용자가 없다면 해당 사용자(④)의 속성 정보를 트위터 REST API에서 불러온다. 트위터를 수집할 때 REST API로 시간 당 불러올 수 있는 트윗 수가 350회로 제한되어 있어 리트윗된 사용자가 누락되기도 한다. 그러나 여기서 리트윗된 사용자는 다른 사용자들에 의해 리트윗이 되기도 하므로 영향력이 어느 정도 있는 사용자로 간주하여 연구에서 제외해서는 안 되는 데이터이다. 그러므로 트위터 DB에 없는 사용자가 생긴다면 REST API에서 해당 사용자의 속성 정보를 따로 불러오도록 한다. 마지막으로 리트윗을 한 사람(③)은 트위터 DB 내에 있는 속성 정보를 추출한다.

분류된 모든 사용자의 속성 정보를 정리하여 필드별로 표 3-9와 같이 정리한다. 이 데이터는 다음 2.3.2에서 가중 네트워크를 형성하는데 입력 데이터로 활용된다.

2.3.2 주제별 트위터 사용자 가중네트워크 생성

본 절에서는 영향력 지수와 주제 유사도를 통한 가중네트워크를 형성한다.

우선 영향력 지수란 트위터 사용자들에게 얼만큼의 영향력을 줄 수 있는가를 의미한다. 영향력 지수는 Leavitt(2009)이 키플레이어는 한 사람의 잠재적 행동이 다른 사람의 행동을 유도하는 사람을 의미한다는 점에서 착안하여 제안한 지표이다. 영향력 지수가 높을수록 다른 사용자들에게 크게 영향을 주며 파급력이 있는 사용자이고, 사용자들이 특정 주제에서 이 사용자를 신뢰하고 따를 확률이 높아진다. 영향력 지수는 다음 식 2-4를 통해 산출할 수 있다.

$$Influence\,Index = \frac{\overline{RT_i} + \overline{Fav_i}}{\overline{RT} + \overline{Fav}} + \frac{\overline{Fol_i}}{\overline{Fol}} \tag{2-4}$$

$$\overline{RT_i} = \frac{\sum_{k=1}^{m} RT_{ik}}{t} \tag{2-5}$$

$$\overline{Fav_i} = \frac{\sum_{k=1}^{m} Fav_{ik}}{t} \tag{2-6}$$

t = 사용자i가주제 내에서 작성한트윗수

$$\overline{RT} = \frac{\sum_{k=1}^{n} \overline{RT_k}}{n} \tag{2-7}$$

$$\overline{Fav} = \frac{\sum_{k=1}^{n} \overline{Fav_k}}{n} \tag{2-8}$$

n = 주제 내의 전체 사용자 수

$$\overline{Fol}_i = \frac{\sum_{k=1}^{m} Fol_{ik}}{t} \tag{2-9}$$

$$\overline{Fol} = \frac{\sum_{k=1}^{n} \overline{Fol}_k}{n} \tag{2-10}$$

여기서 $\overline{RT_i}$는 사용자 i가 주제 내에서 쓴 트윗에 해당하는 리트윗 횟수를 모두 더해 주제 내에서 쓴 트윗 수로 나눈 사용자 i의 평균 리트윗 수이다. $\overline{Fav_i}$는 $\overline{RT_i}$와 같이 사용자 i의 평균 favorite 수를 나타낸다. 그리고 $\overline{RT}$와 $\overline{Fav}$는 주제 내의 전체 사용자들의 평균 리트윗과 평균 favorite 수를 나타낸다. 리트윗과 favorite의 경우 사용자가 쓴 텍스트별로 값이 모두 다르므로 평균값을 구하였다. 사용자별 리트윗과 favorite 수를 전체 평균으로 나누는 것은 해당 사용자의 텍스트에 대한 가중치를 의미하는 것으로, 사용자가 쓴 텍스트가 전체 평균값보다 리트윗과 favorite이 많이 되면 될수록 그 주제에서 영향력은 커진다. 다음으로 $\overline{Fol_i}$는 앞서 $\overline{Fav_i}$과 $\overline{RT_i}$를 구한 것과 같이 사용자 i가 주제 내에서 쓴 트윗에 해당하는 팔로워 수를 트윗 수로 나눈 평균값을 의미한다. 그리고 $\overline{Fol}$는 전체 사용자들의 평균 팔로워 수를 전체 사용자의 수로 나눈 평균값을 의미한다. 전체 사용자의 평균보다 사용자 i의 팔로워 수가 높으면 높을수록 다른 트위터 사용자들이 이 사용자를 믿고 따른다는 의미로 볼 수 있다. 결과적으로

영향력 지수는 한 사용자가 작성한 텍스트를 다른 사용자들이 전달하며 널리 퍼뜨리고 있다는 지표인 리트윗과 다른 사용자들이 좋아요를 누르며 그 텍스트에 대한 호감도가 있다는 것을 알려주는 favorite, 그리고 그 사용자를 따르는 다른 사용자들의 수를 알 수 있는 팔로워 수를 더한 비율 값을 구함으로 이 사용자가 얼마나 영향력을 끼치는지 알려주는 지표이다. 영향력 지수가 높아지면 높아질수록 다른 사용자가 해당 사용자의 계정에 방문할 확률이 높아진다.

다음으로 주제 유사도란 특정 주제에서 사람들 간의 텍스트가 얼마나 유사한가를 구하는 것을 의미한다. McPherson(2001)에 의하면 유사성은 연결을 만들어낸다. 왜냐하면, 사람들은 자신과 비슷한 사람들과 교류하려고 하기 때문이다. 보다 구조적으로 비슷한 사람끼리 어떠한 이슈에 관련한 커뮤니케이션을 활발히 하고 서로의 이슈 직책에 참여할 가능성이 크며, 이는 서로에게 더 많은 영향을 미친다. 따라서 특정 주제에 관하여 사용자들 간의 주제 유사도가 높을수록 서로의 관심사가 비슷하다고 할 수 있다. 주제 유사도를 구하기 위하여 유사도를 구하는 방법의 하나인 Jaccard Index를 사용하였다. Jaccard Index란 두 집합 사이의 유사도를 측정하는 방법으로 0과 1 사이의 값을 가지며 동일한 원소가 한 개도 존재하지 않으면 0 값을 가지고 모두 일치하면 1 값을 가진다. Jaccard Index는 식 2-11, 2-12와 같다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2-11}$$

$$0 \leq J(A,B) \leq 1 \tag{2-12}$$

본 연구에서는 키워드를 포함하는 텍스트 간의 유사도를 구하기 때문에 Jaccard Index를 이용하였다. 결과적으로 사용자 간의 텍스트에서 공통으로 등장하는 명사들을 멀티 셋으로 만든 후 Jaccard Index를 통해 전체 명사들의 멀티 셋에서 공통되는 명사가 어느 정도 나오는지를 통해 주제 유사도를 구할 수 있다. 주제 유사도를 통해 사용자 간의 유사한 정도를 확인할 수 있고, 유사할수록 같은 주제에 관심을 가지는 사용자라고 볼 수 있다.

2.4 KeyplayerRank를 이용한 키플레이어 탐지

본 연구에서 이용하는 KpRank는 구글의 페이지랭크를 변형시킨 것이다. 페이지랭크란 웹 페이지의 상대적 중요성을 측정하기 위해 웹 그래프를 기반으로 웹 페이지를 랭킹(ranking)하는 방식을 말한다(Brin & Page, 1998). 페이지랭크는 검색엔진에서 사용자가 찾고자 하는 관련된 페이지만을 찾아볼 수 있도록 검색엔진의 품질을 향상시키고자 개발된 알고리즘이다. 이 알고리즘은 학술지 인용 방식에 착안하였지만, 특정 웹 페이지를 인용하는 다른 웹 페이지들이 얼마나 많이 있느냐를 생각하는 대신 그 웹 페이지에 걸린 링크 숫자를 정규화하는 방식을 사용하였다. 따라서 중요한 웹 페이지를 인용하는 한 웹 페이지의 랭크는 다른 웹 페이지에 비해 높아진다는 것이다.

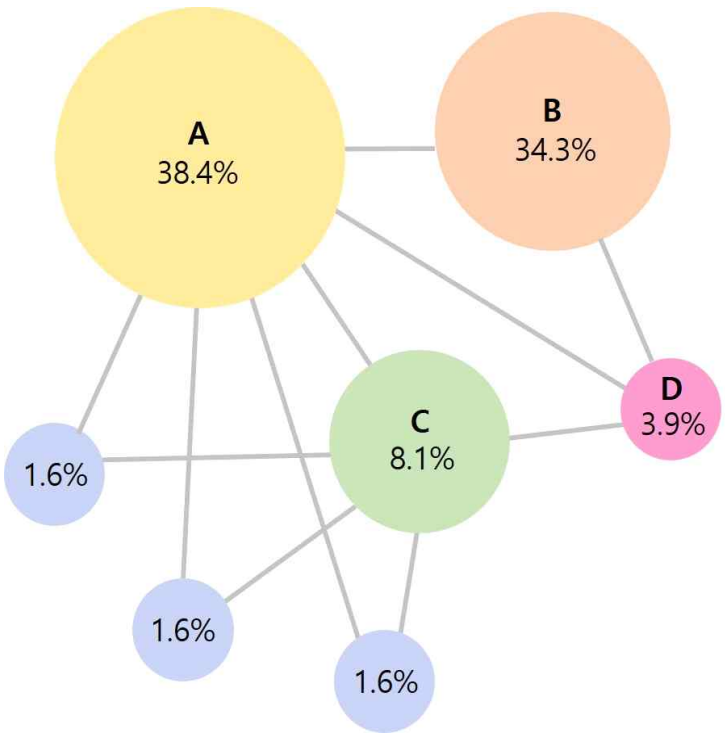

그림 2-5. 페이지랭크 알고리즘 그래프

그림 2-5에서 각 퍼센트는 다른 웹 페이지에서 이 페이지로 올 확률을 의미한다. ‘B’는 다른 웹페이지들과 비교하여 연결된 링크 수가 적으나, 가장 영향력 있는 ‘A’ 페이지와 연결되면서 상대적인 중요도가 올라간 것을 확인할 수 있다.

페이지랭크 식은 식 2-13과 같다.

$$PR(p_i) = \frac{(1-d)}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad (2\text{-}13)$$

여기서, N은 모든 페이지의 수이고 $M(p_i)$는 p_i로 향하는 페이지들의 군집을 말한다. $L(p_j)$는 p_j와 이어지는 모든 링크 수를 의미한다. 또한, 여기서 d는 damping factor라 하여 웹 서핑을 하는 사람이 한 페이지에 만족하지 못하고 다른 페이지의 링크로 갈 확률을 의미하며, 일반적으로 0.85의 값을 쓴다. 정리하면 p_i의 페이지랭크 값은 p_i와 연결된 모든 다른 페이지들의 페이지랭크 값을 정규화시킨 값의 합이라고 할 수 있다. 따라서 모든 페이지의 페이지랭크 값을 더하면 1과 같다.

본 연구에서는 영향력 지수와 주제 유사도를 모두 적용하는 KpRank를 정의하였으며, 이는 식 2-14와 같다.

$$KpRank(p_i) = (1-d) \bullet Influence\,Index + d \sum KpRank(p_j) \bullet Topical\,Similarity \quad (2\text{-}14)$$

여기서 d는 페이지랭크와 유사하지만 본 연구에서는 특정 주제 내의 네트워크이기 때문에 트위터 사용자가 관심 있는 주제 내를 돌아다닐 확률을 의미한다. 페이지랭크와 달리 n으로 나누지 않는데, 그 이유는 영향력 지수에서 전체 사용자들의 가중치 값으로 나눠주기 때문이다.

2.3장에서 영향력 지수와 주제 유사도를 구하여 식 2-14에 넣어주었다. 여기서 영향력 지수는 사용자 p_i의 트위터를 다른 사용자들이 방문할 확률을 높여주는 가중치로 이용되고, 주제 유사도는 같은 주제에 관심을 보이는 사용자의 트윗을 방문할 확률을 높여준다. 따라서 트위터 사용자 p_i의 KpRank는 사용자 p_i와 연결된 같은 주제를 언급하는 다른 사용자들의 KpRank의 합산과 사용자 p_i의 영향력이 가중되어 다른 사용자들이 사용자 p_i의 트위터에 머물 확률을 더한 값이다. 그러므로 KpRank는 특정 주제 내에서 상대적으로 영향력 있는 트위터 사용자를 찾는 알고리즘이다. 특정 주제 내에서 상대적으로 영향력 있는 사용자를 찾는 것은 그 사용자가 주제에서 전문적인 정보와 지식을 가지고 있으며 가장 영향력 있는 사용자이기 때문이다. KpRank를 통해 가장 높은 랭크를 가지는 사용자를 키플레이어라고 부르며, 이 사용자가 해당 주제 내에서 가장 영향력 있는 사용자라고 판단한다.

3. 실험 적용 및 결과

3.1 데이터 수집 및 전처리

1) 조선일보 데이터

본 연구에서 이용하는 조선일보 데이터는 웹 스크래퍼(web scraper)를 이용하여 뉴스 기사를 웹 스크래핑(web scraping)[6]하여 수집했다. 웹 스크래퍼에는 기사의 키워드, 시작 날짜와 마지막 날짜, 게시된 기사의 시작 번호, 스크랩한 데이터 결과 저장 파일명을 입력하여 웹 스크래핑한다(표 3-1).

표 3-1. 조선일보 웹 스크래퍼에 들어가는 옵션 내용

옵션	내용
keyword	검색 키워드 입력
date_from	게시된 기사의 시작 날짜 ex)yyyymmdd
date_to	게시된 기사의 마지막 날짜 ex)yyyymmdd
start	게시된 기사의 시작 번호 (>=1)
result_file	검색 결과 저장 파일명

웹 스크래퍼를 이용하여 수집한 데이터는 tsv형식[7]으로 저장되며, Microsoft Excel을 이용하여 데이터를 열어보면 그림 3-1과 같다. 수집된 데이터는 'TITLE, CATEGORY, URL, DATE, TEXT' 필드로 이루어져 있다(표 3-2).

---

6) 인터넷 웹 페이지에 나타나는 데이터 중에서 필요한 데이터만을 추출하도록 만들어진 프로그램이며, 각 사이트로부터 데이터를 수집해 오는 기술, 일정 포맷으로 변환하는 기술을 말함(한국정보통신기술협회 정보통신용어사전).
7) 열 구분 기호가 탭 문자이고 레코드 구분 기호가 줄 바꿈 문자인 위치의 쉼표로 구분된 데이터 형식.

표 3-2. 조선일보 데이터 필드 내용

필드명	필드 내용
TITLE	기사 제목
CATEGORY	기사의 범주
URL	기사의 웹 주소
DATE	기사가 작성된 날짜
TEXT	기사의 내용

TITLE	CATEGORY	URL	DATE	TEXT
혼다·소프트뱅크 손잡고 말하는 지	자동차	http://biz.chosun.com/site/	2016-07-21	일본 자동차 기업 혼다가 IT·통신 기업 소프트뱅크
로봇의 넘어짐에서 배운다. 인간도	뉴스	http://biz.chosun.com/site/	2016-07-21	"로봇이 넘어지고 고장나지 않으면 아무 것도 배울
테슬라 "사람보다 10배 안전운전하	자동차	http://biz.chosun.com/site/	2016-07-21	일론 머스크 테슬라 최고경영자(CEO)가 테슬라 블
경제적 행복 막는 장애물 1위는 '노	뉴스	http://biz.chosun.com/site/	2016-07-21	현대硏, 성인남녀 1012명 설문조사...56.2% "하반기
SK C&C, 스마트 팩토리 솔루션 '스	뉴스	http://biz.chosun.com/site/	2016-07-21	SK C&C는 스마트 팩토리 종합 솔루션 브랜드 '스
[비즈 발언대] 제4차 산업혁명 시대	뉴스	http://biz.chosun.com/site/	2016-07-21	"한국은 지금 선택의 기로에 있다." 최근 별세한 미
'대뇌 지도' 첫 완성... 치매 등 연구	뉴스	http://biz.chosun.com/site/	2016-07-21	[오늘의 세상]<NL>기능별로 180개 영역 구분 "뇌
국내 최초 코딩 교육 의무화... SW	사회	http://news.chosun.com/sit	2016-07-21	세종대학교<NL>세종대학교는 내년 3월 '소프트웨
치매 잡는 뇌연구 신기원 열었다...	뉴스	http://biz.chosun.com/site/	2016-07-21	인간 문명 발달의 근원인 뇌의 작동 원리를 파악하
허창수 GS 회장 "10년 후 내다보고	뉴스	http://biz.chosun.com/site/	2016-07-20	"시장 변화의 맥을 잘 잡아 5년, 10년 후를 내다보
[팝콘뉴스] 알파고, '세계 최강 바둑	소년조선	http://kid.chosun.com/site/	2016-07-20	세계 바둑랭킹 1위 자리에 새로운 이름이 올랐습니
카카오, 주문중개 플랫폼 기업 '씨(	뉴스	http://biz.chosun.com/site/	2016-07-19	카카오(035720)는 19일 주문중개 플랫폼 기업 '씨(
[특징주]日소프트뱅크 ARM 인수..	스포츠 · 연예	http://news.chosun.com/sit	2016-07-19	일본의 IT·통신기업 소프트뱅크가 영국의 반도체 설
IBM, "왓슨 덕에 예상보다 높은 2(	뉴스	http://biz.chosun.com/site/	2016-07-19	미국 IBM이 2분기에도 매출이 감소하며 17분기 연
알파고, 세계 바둑 랭킹 1위 등극...	스포츠 · 연예	http://news.chosun.com/sit	2016-07-19	구글의 인공지능 바둑 프로그램 알파고(AlphaGo)

그림 3-1. 웹 스크래핑을 통한 조선일보 데이터

본 연구에서는 2015년 7월 22일부터 2016년 7월 22일까지 약 1년간의 조선일보 기사를 수집하였다. 키워드 '인공지능, 자율주행, 드론'으로 1년간 수집한 데이터는 총 8,890개이다. 수집한 조선일보 기사에서 주제를 추출하기 위해 기사 내용이 담긴 'TEXT' 필드만을 전처리하였다. 기사 텍스트를 처리하기 위하여 본 연구에서 사용하는 프로그래밍 언어는 Python 3.6.2이며, 한글로 이루어진 자연어처리를 위해 KoNLPy(Korean Natural Language Processing in Python) 패키지를 이용하였다.

기사 텍스트는 여러 문장으로 구성되어 있으므로 먼저 문장 단위로 분리하는 과정을 진행하였다. 이러한 과정을 토큰화(Tokenization)라고 한다. 토큰화는 문서나 문장을 분석하기 좋도록 토큰(token)[8]으로 나누

8) 토큰(token)이란 의미를 가지는 문자열을 뜻하며, 형태소나 단어를 포함함.

는 작업을 말한다. 문장을 분리한 후 문장 내에서 형태소 분석을 하기 위해 단어 단위로 한 번 더 토큰화하였다. 문장 및 단어 단위로 토큰화한 텍스트에서 텍스트 분석에 의미가 없다고 판단되는 불용어를 제거하였다. 이때, 제거된 불용어는 한자, 숫자, 영문자, URL 주소이다. 불용어가 제거된 텍스트에서 주제를 추출하기 위해 명사만을 이용하였다. 명사만을 추출하기 위해 어절 단위로 분리된 단어에 품사(POS, part-of-speech)를 부여하는 형태소 분석을 시행하였다. 여기서 이용한 형태소 분석기는 java기반의 한국어 형태소 분석기로, 각 단어에 대해 42종의 품사를 부여하는 KoNLPy 패키지의 KOMORAN을 이용하여 명사(일반명사(NNG), 고유명사(NNP))만을 추출하였다. 그림 3-2의 예시를 보면 문장에 NNP, JKS, NNG, JKO, VV, EP, EC와 같은 품사가 태그되었다. NNP는 고유명사, JKS는 주격조사, NNG는 일반명사, JKO는 목적격 조사, VV는 동사, EP는 선어말어미, EC는 연결어미를 의미한다. 품사를 태깅한 후 NNP와 NNG만을 추출하였다.

그림 3-2. 형태소 분석의 예시

명사만 추출한 조선일보 기사 텍스트를 저장하여 다음 과정인 SVM 텍스트 분류에 이용하였다.

2) 트위터 데이터

본 연구에서 사용되는 트위터 데이터는 REST 기반의 오픈 API를 통해 수집하였다. 조선일보 수집 기간과 같이 2015년 7월 22일부터 2016년 7월 22일까지 약 1년간의 데이터를 수집하였으며, 수집된 데이터를 중복 제거하여 총 3,001,589개의 트위터 데이터를 이용하였다. 트위터 오픈 API에서 제공하는 다양한 속성 정보 중 본 연구에서는 text, user_id, user_s_name, followers_count, retweet_count, favorite_count 필드만을 이용하였다(표 3-3).

표 3-3. 본 연구에서 사용한 트위터 속성 정보 필드 내용

필드명	타입	설명
text	문자	UTF-8 표준 언어 코드로 작성된 상태 업데이트에 대한 전체 내용
user_id	정수	해당 트윗에 대한 고유 식별자의 정수 표현
user_screen_name	문자	사용자가 자신을 식별하는 화면 이름 또는 별칭
followers_count	정수	해당 계정의 현재 팔로워 수
retweet_count	정수	사용자가 쓴 해당 트윗이 리트윗된 횟수
favorite_count	정수	트위터 사용자들이 해당 트윗을 얼마나 좋아했는지를 표현한 횟수

트위터 필드 중 트위터 텍스트 분석을 하기 위하여 'text' 필드만을 전처리하였다. 전처리 과정은 앞서 조선일보 텍스트를 전처리했던 것과 같이 KoNLPy 패키지를 이용하여 불용어 제거 및 형태소 분석을 수행하였다. 트위터에서는 SVM을 통한 텍스트 분류와 KpRank를 구하기 위한 주제 유사도를 계산하기 위하여 명사만을 추출하였다.

3.2 결과 및 토의

3.2.1 SVM 결과

1) 조선일보 데이터

본 연구에서는 연관성 있는 신문기사를 가지고 주제를 추출하기 위해 선형 SVM을 통한 문서 분류를 수행하였다. SVM 과정은 다음과 같이 이루어진다. 분석에 이용될 특징을 선정하고, 특징 가중치로 특징을 표현하는 텍스트 인덱싱(Indexing)을 한 뒤, 데이터를 3:1의 비율로 트레이닝 데이터(training data)와 테스트 데이터(test data)로 나누었다. 그리고 트레이닝 데이터를 통해 SVM을 훈련하고, 훈련된 SVM 모델을 통해 텍스트를 분류하는 과정으로 진행된다.

SVM을 하기 위해 이용된 프로그래밍 언어는 R x64 3.4.1이며, Document Term Matrix(DTM)[9]를 생성하기 위한 'tm', SVM 모델을 생성하는데 필요한 'e1071', confusion matrix를 만들기 위한 'caret' 패키지를 이용하였다.

먼저 앞서 명사만을 추출한 조선일보 텍스트들에 라벨링을 수행하였다. 예를 들어, 인공지능과 관련이 있는 텍스트의 class에는 '1'을 부여하고 관련이 없는 내용을 가진 텍스트에는 '-1' 값을 class에 부여하였다(표 3-4).

9) 문서 군에서 발생하는 용어의 빈도를 설명하는 수학적 행렬.

표 3-4. 조선일보 텍스트 라벨링 과정

No.	text	class
1	[‘일본’, ‘자동차’, ‘혼다’, ‘통신’, ‘기업’, ‘소프트뱅크’, ‘인공지능’, ‘이용’, ‘자동차’]	1
2	[‘주문’, ‘플랫폼’, ‘기업’, ‘지분’, ‘취득’, ‘국내’, ‘대표’, ‘프랜차이즈’, ‘브랜드’, ‘중개’, ‘벤처기업’]	-1
3	[‘구글’, ‘인공지능’, ‘바둑’, ‘프로그램’, ‘알파’, ‘중국’, ‘커제’, ‘사이트’, ‘집계’, ‘순위’, ‘알파’, ‘커제’, ‘커제’, ‘해협’, ‘바둑’, ‘스웨’, ‘이세돌’]	1
4	[‘사드’, ‘고고’, ‘미사일’, ‘방어’, ‘체계’, ‘잠수함’, ‘발사’, ‘탄도미사일’, ‘대한민국’, ‘영토’, ‘사드’, ‘배치’, ‘중국’, ‘경제’, ‘보복’]	-1
5	[‘서울’, ‘한복판’, ‘광화문’, ‘일대’, ‘도심’, ‘풍경’, ‘빌딩’, ‘주변’, ‘고층’, ‘빌딩’, ‘시야’, ‘경우’, ‘직원’, ‘출입’, ‘오피스’]	-1

라벨링 한 5,000개의 텍스트를 하나의 말뭉치(corpus)로 형성하였다. 텍스트 분류를 하는 데 있어 유용하게 사용할 만한 단어들을 선정하기 위하여 특징 선정과정으로 문헌 빈도(Document Frequency)를 사용하였다. 이때, 문헌빈도가 2 이하인 단어들을 제외하고 나머지 단어들을 특징으로 선정하였다.

다음 과정인 텍스트 인덱싱에서는 특징 가중치로 Term Frequency - Inverse Document Frequency(TF-IDF)[10]를 적용하였다. DTM에 특징 가중치인 TF-IDF를 적용하면 그림 3-3과 같다.

10) 정보 검색과 텍스트 마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치.

	감정	개발	경험	공동	기반	기분	기술	기업	대화	도쿄	로봇
1	0.15809043	0.036691940	0.034042934	0.036869995	0.033171089	0.05384137	0.040696648	0.068840750	0.096151813	0.060043234	0.189596560
2	0.00000000	0.019353991	0.017956712	0.000000000	0.000000000	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.400027907
3	0.00000000	0.039138070	0.000000000	0.000000000	0.000000000	0.00000000	0.072349596	0.000000000	0.000000000	0.000000000	0.000000000
4	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.00000000	0.005366591	0.000000000	0.000000000	0.000000000	0.000000000
5	0.00000000	0.000000000	0.000000000	0.000000000	0.050148418	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
6	0.00000000	0.000000000	0.000000000	0.000000000	0.015092060	0.00000000	0.018516010	0.020880607	0.000000000	0.000000000	0.017252388
7	0.00000000	0.000000000	0.000000000	0.020698945	0.000000000	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
8	0.00000000	0.000000000	0.000000000	0.010597364	0.009534205	0.00000000	0.040940340	0.000000000	0.000000000	0.000000000	0.000000000
9	0.00000000	0.005755598	0.000000000	0.000000000	0.000000000	0.00000000	0.006383788	0.000000000	0.000000000	0.000000000	0.011896255
10	0.00000000	0.022015164	0.000000000	0.000000000	0.000000000	0.00000000	0.012208994	0.000000000	0.000000000	0.000000000	0.000000000
11	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
12	0.00000000	0.068618694	0.000000000	0.000000000	0.000000000	0.00000000	0.076108017	0.028609143	0.000000000	0.000000000	0.047276025
13	0.00000000	0.015585957	0.000000000	0.000000000	0.028180748	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
14	0.00000000	0.000000000	0.000000000	0.000000000	0.024308585	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
15	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.00000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000

그림 3-3. TF-IDF를 특징 가중치로 적용한 조선일보 텍스트 DTM

위 그림의 DTM에서 각 행은 텍스트를 나타내고, 각 열은 이전에 선정된 특징을 의미한다. 그리고 행렬 안의 각 요소는 TF-IDF가 가중된 값을 의미한다.

DTM을 가지고 SVM 모델을 형성하기 위한 트레이닝 데이터와 테스트 데이터를 3:1의 비율로 나누었으며, 트레이닝 데이터는 3,750개이고, 테스트 데이터는 1,250개였다. 트레이닝 데이터를 가지고 최적의 학습 조건을 맞추기 위하여 파라미터인 cost와 gamma를 조정하였다. 적절한 파라미터 값을 정하기 위하여 10겹 교차 타당성 평가(10-fold cross-validation)를 진행하였다. cost 값으로 0.1, 1, 10과 gamma 값으로 0.1, 1, 10의 총 9개의 조합 중 최적의 조합을 찾아내기 위해 교차 타당성 평가를 수행하였고, 그 결과 cost 값으로 10과 gamma 값으로 0.1이 최적의 조합으로 결정되어 이를 트레이닝 데이터로 SVM을 훈련하는 데 이용하였다.

훈련된 SVM 모델로 1,250개의 테스트 데이터에 적용하여 분류하고, 분류가 정확하게 됐는지를 검증하기 위해 표 3-5의 값을 식 3-1, 3-2, 3-3을 통해 정확률(precision)과 재현율(recall), F1-measure를 구하였다(표 3-7).

표 3-5. Confusion Matrix

		Actual data	
		1	-1
Predicted data	1	True Positive(TP)	False Positive(FP)
	-1	False Negative(FN)	True Negative(TN)

표 3-6. 조선일보 test data Confusion matrix

	Actual			
Predicted		1	-1	sum
	1	200	76	276
	-1	91	883	974
	sum	291	959	1,250

$$\text{정확률}(p) = \frac{TP}{TP+FP} \tag{3-1}$$

$$\text{재현율}(r) = \frac{TP}{TP+FN} \tag{3-2}$$

$$F_1 - measure = 2 \cdot \frac{p \cdot r}{p+r} \tag{3-3}$$

표 3-7. 조선일보 SVM 평가

모델	Precision	Recall	Accuracy	F1-measure
SVM(조선일보)	0.7246	0.6873	0.8664	0.7055

테스트 데이터 1,250개에 대해 검증을 한 결과, 표 3-6에서 Actual은 라벨링 과정을 통해 1과 -1 값을 부여한 것이고, Predicted는 SVM 모델을 통해 분석한 결과이다. 1,250개의 데이터 중 올바르게 분류된 것은

1,083개로, 약 86.64%의 정확도(overall accuracy)를 보였다. 그리고 정확률은 약 72.46%로 높은 예측력을 보였고, 실제 관련 있는 텍스트 중 올바르게 예측된 텍스트의 비율인 재현율은 약 68.73%였다. 라벨링 한 값과 SVM 모델을 통해 예측된 값이 모두 관련 있는 것으로 나온 200개의 데이터에 대해 다음 과정인 LDA를 진행하였다.

2) 트위터 데이터

트위터 SVM 과정은 LDA를 통해 추출된 주제별 키워드를 포함하는 트위터들을 전처리한 후 진행된다. 3.2.2에서 나온 LDA 결과 중 주제 1번의 키워드 '이세돌, 바둑, 알파고, 인공지능, 대국'을 이용하여 추출한 총 트위터 5,666개를 가지고 SVM을 수행하였다. 트위터에서도 SVM을 수행하는 이유는 조선일보 텍스트를 분류한 것과 같이 트위터에서도 주제와 관련 있는 텍스트를 쓴 사용자를 뽑기 위함이다.

앞서 조선일보 텍스트에 SVM을 한 것과 같은 과정을 진행하였으며, 트레이닝 데이터와 테스트 데이터만 달라졌다. 트위터에서 이용한 트레이닝 데이터는 조선일보 데이터 5,000개와 임의로 뽑은 트위터 데이터 400개로 구성되어 있다. 조선일보 데이터만 훈련한 모델을 트위터에 적용하기에는 DTM 형성에 적합하지 않아 트위터 데이터를 추가하여 트레이닝 데이터로 만들었다. 그리고 트위터 5,666개 중 형태소 분석결과 null 값이 나온 67개의 데이터를 제외한 5,599개의 데이터를 테스트 데이터로 정하였다.

트위터 SVM을 돌리는데 이용한 파라미터 값은 이전에 10겹 교차 타당성 평가로 얻어진 파라미터를 이용하였다. 분류가 정확하게 됐는지 검증하기 위하여 정확률, 재현율 그리고 F1-measure를 구하였다(표 3-9).

표 3-8. 트위터 test data Confusion matrix

		Actual		
Predicted		1	-1	sum
	1	301	316	617
	-1	6	4,976	4,982
	sum	307	5,292	5,599

표 3-9. 트위터 SVM 평가

모델	Precision	Recall	Accuracy	F1-measure
SVM(트위터)	0.4878	0.9805	0.9425	0.6515

전체 5,599개 중 5,277개의 값이 올바르게 분류되었고, 약 94.25%의 높은 정확도를 가진다. 그에 반해 예측된 텍스트 중 관련 있다고 나온 비율은 약 48.78%로 낮은 정확률을 보여주고 있다. 그러나 실제로 관련 있는 텍스트 중 올바르게 예측된 비율은 약 98.05%로 높은 재현율을 보였다. 표 3-8 결과를 보면 관련 있다고 분류된 데이터의 개수가 현저히 낮은 것을 확인할 수 있다. 이것은 트위터 텍스트에서 주제의 키워드 중 '대국'이 들어간 텍스트를 추출하였을 때, '여기 순대국이 정말 맛있어요!!!', 'I am at 할매순대국', '대국남아 파이팅!', '쨔쟈쟌(@ 통큰원조할매토종순대국 in 노원구, 서울특별시, 서울특별시, 서울)' 등 대국이 앞뒤로 들어간 텍스트를 추출함으로 관련 없다고 라벨링 된 개수가 관련 있다고 라벨링 된 개수보다 상대적으로 많았기 때문이다. 그런데도 SVM 모델을 통해 관련 있다고 나온 데이터 수가 라벨링 된 개수보다 두 배로 많이 예측되었기 때문에 낮은 정확률을 보였다. 이는 키워드 앞뒤에 불필요한 글자를 제거함으로써 개선해야 할 필요가 있다.

따라서 라벨링 값과 SVM 모델을 통해 예측된 값이 모두 관련 있다고 나온 301개의 데이터에 대해 주제별 사용자 분류를 수행하고 네트워크를 형성하였다.

3.2.2 LDA 결과

3.2.1에서 조선일보 텍스트에서 명사만 추출하여 SVM으로 '인공지능'과 관련 있다고 나온 200개의 데이터를 가지고 LDA를 수행하였다. LDA는 Python 3.6.2 프로그래밍 언어를 이용하여 실험하였고, 토픽모델링 패키지인 gensim을 사용하였다.

LDA 모델을 만들기 위해서 먼저 명사로 이루어진 200개의 텍스트를 DTM으로 생성하였다. 각 단어가 텍스트당 몇 번의 빈도로 언급되었는지 DTM으로 나타낸 뒤, 각 단어 토큰에 고유한 아이디를 부여하였다. 그리고 전체 텍스트들을 하나의 말뭉치로 만들었다. 말뭉치로 LDA 모델을 생성하였고, LDA 모델을 생성하는데 들어간 파라미터는 num_topics와 passes이었다. num_topics는 잠재적인 주제의 개수를 정하는 것이고, passes는 모델이 말뭉치를 통과하는 횟수를 정하는 것으로, passes의 횟수가 많아질수록 모델이 정교해진다. 본 연구에서는 num_topics를 10으로 정하고 passes는 1,000으로 설정하여 LDA 모델을 생성하였다(표 3-10).

표 3-10. ‘인공지능’ LDA 결과

주제 번호	키워드 1	키워드 2	키워드 3	키워드 4	키워드 5
1	0.035 * “이세돌”	0.035 * “바둑”	0.033 * “알파고”	0.028 * “인공지능”	0.024 * “대국”
2	0.016 * “데이터”	0.015 * “분석”	0.011 * “개발”	0.010 * “빅데이터”	0.008 * “기술”
3	0.033 * “자율”	0.019 * “기술”	0.018 * “자동차”	0.016 * “로봇”	0.015 * “구글”
4	0.014 * “자율”	0.014 * “자동차”	0.014 * “업체”	0.010 * “협력”	0.010 * “포드”
5	0.019 * “인공지능”	0.015 * “인간”	0.011 * “인텔”	0.007 * “무인”	0.007 * “기술”
6	0.017 * “구글”	0.017 * “데이터”	0.017 * “왓슨”	0.015 * “서비스”	0.013 * “기술
7	0.017 * “클라우드”	0.012 * “기계”	0.011 * “인간”	0.010 * “기술”	0.008 * “스마트”
8	0.104 * “드론”	0.015 * “비행”	0.011 * “조종”	0.009 * “규제”	0.008 * “사람”
9	0.015 * “기술”	0.012 * “인공지능”	0.011 * “로봇”	0.011 * “개발”	0.010 * “지능”
10	0.048* “인공지능”	0.043 * “로봇”	0.014 * “인간”	0.013 * “연구”	0.013 * “기술”

표 3-10을 보면 주제 번호가 나오고 그 주제에 속하는 5개의 키워드와 각 키워드별 확률값이 함께 나온다. 이 확률값은 해당 주제에서 해당 키워드가 등장할 확률이다.

LDA 결과 '인공지능' 단어가 주제별로 많이 검출되었고, 인공지능과 관련한 '로봇', '기술', '기계' 등의 키워드가 주제에 고루 분포된 것을 확인할 수 있다.

주제 1번으로 키워드 '이세돌, 바둑, 알파고, 인공지능, 대국'이 나왔다. 이 키워드들로 유추해 봤을 때 주제 1번은 '알파고와 이세돌의 바둑대국'이라 할 수 있다. 2016년 3월 9일부터 15일까지 알파고와 이세돌의 바둑 대결이 인공지능과 인간의 대결로 한창 이슈가 되고 주목을 받았었다. 따라서 본 연구의 수집 기간인 2015년 7월 22일부터 2016년 7월 22일 기간에 이와 관련된 뉴스 기사도 많이 발행되었기 때문에 1번과 같은 주제가 나왔을 것으로 추정된다.

이외에도 2번 주제에서 '빅데이터'와 관련된 '데이터', '분석' 등의 키워드들이 추출되었고, 3번에서는 구글의 자율주행 자동차 관련 키워드가 뽑혔다. 4번 주제에서는 키워드로 자율주행 자동차 선도 업체인 '포드'가 뽑혔다. 실제로 포드는 2017년 1월에 신형 자율주행 자동차를 선보인 적이 있다. 또한, 주제 8번의 키워드들을 보면 드론 비행 규제를 의미하고 있다. 2016년 3월에 항공안전법이 제정되면서 이와 관련한 뉴스 기사도 많이 발행되어 8번과 같은 주제가 추출된 것을 알 수 있다.

3.2.3절에서는 LDA 결과 중 1번 주제인 '알파고와 이세돌의 바둑대국'을 가지고 트위터에서 이 주제에 해당하는 사용자들 간의 네트워크를 형성하였다.

3.2.3 사용자 간 네트워크 생성 결과

‘알파고와 이세돌의 바둑대국’과 관련된 주제에 해당하는 트위터들을 추출한 결과 총 3,001,589개의 트위터 데이터 중 6,266개의 트위터가 추출되었다. 6,266개의 데이터에서 중복된 데이터를 제거한 결과 총 5,666개의 트위터를 얻었다. 그리고 주제와 관련 있는 트위터를 선별하기 위해 3.2.1절에서 SVM을 수행하여 얻은 301개의 데이터에서 사용자 간 네트워크를 형성하였다. 301개의 데이터는 사용자가 작성한 text, 사용자 아이디별 고유의 숫자인 user_id, 사용자가 지정한 별칭인 user_name, 사용자의 아이디인 user_screen_name, 해당 사용자를 팔로우하는 사용자의 수를 나타내는 followers_count, 해당 사용자가 쓴 내용이 리트윗 내용일 경우 해당 리트윗된 사용자의 리트윗된 횟수를 나타내는 RT_count, 그리고 사용자가 쓴 트위터를 ‘좋아요’ 한 횟수를 나타내는 Fav_count(favorite_count)로 이루어져 있다(표 3-11).

추출된 트위터 데이터들의 text를 보면, 다른 사용자의 내용을 리트윗한 text가 201개로 사용자가 직접 작성한 text 100개보다 약 2배가량 많았다. 트위터 특성상 다른 사용자의 글을 퍼다 나르는 리트윗이 빈번하게 이루어지기 때문에 상대적으로 다른 사용자의 내용을 담은 리트윗된 text가 많았다. 리트윗한 텍스트는 RT_count가 리트윗된 사용자의 리트윗 횟수이기 때문에 리트윗된 사용자를 따로 분류하였다(그림 2-4).

먼저 리트윗을 한 사람과 안 한 사람을 분류하였다. 총 사용자 172명 중 리트윗을 안 한 사용자는 45명, 리트윗한 사용자는 127명이었다. 다음으로 리트윗을 한 사용자의 텍스트에 있는 리트윗된 사용자의 아이디를 확인하였다. 그 결과 사용자 11명이 수집한 트위터 DB 내에 존재하지 않았다. 그 이유는 REST API로 트위터를 수집할 때 트위터를 가져오는 개수가 한정되어 있어 누락된 사용자가 발생했기 때문이다. 따라서, 리트윗한 사용자의 텍스트에서 ‘RT @’뒤의 screen name을 뽑아 따로 저장하였다. 리트윗된 사용자 중 트위터 DB 내에 없는 사용자는 다음 표 3-12와 같다. 이때, 사용자의 팔로워 수는 DB 내에 존재하지 않아 알 수 없으므로 트위터 API를 통해 각 사용자의 팔로워 수를 수집하였다.

표 3-11. '알파고와 이세돌의 바둑대국'에 해당하는 트위터 데이터 일부

No	text	user_id	user_name	user_screen_name	followers_count	RT_count	Fav_count
1	[바둑] 강경한 이세돌 "바둑계, 상식이 통하게 만들어야한다" https://t.co/3N3619Sw7R	916341679	블**	jhans2**	30	0	1
2	RT@iamkepco: 어제 내린 비로 조금 추워진 출근길... 오늘은 이세돌 대 알파고의 바둑 첫 대결이 있는 날이네요.	525546227	kremi**	minsu22**	673	2	0
3	지금 바둑 상황이 완전 진흙탕 싸움이 된건가??	309482925	Dev.Si**	rainloop**	262	1	0
4	인공지능 알파고의 첫승! 경우의 수가 너무 많아 기계가 인간을 이길 수 없다고	525062957	한국전력** (KEPCO)	iamkep**	23,986	2	0
5	이번 이세돌 VS 알파고는 바둑에서 이세돌이 다른 사람이 넘어설 수없는 인간의 정점에있기 때문 아닌가.....	178693143	작은 **	poca_preghie**	5	0	0

표 3-12. 트위터 DB 내에 없는 리트윗된 사용자들의 속성 정보

user_s_name	user_id	user_name	followers_count	RT_count	Fav_count
romnb**	472494203	배*숙(Bae *Sook)	2,426	1	0
ooobbb**	4121863584	은혜 (산성동)	3,931	26	0
wikitr**	47856818	위키*리 WIKITR**	418,245	80	0
kbs01**	3671291773	KYL*	487	3,924	0
yonhaptv_**	3311067703	연합** Entertainment	340	6	0
HuffPostKor**	543778633	허프포스트코**	821,127	51	0
sisain_edit**	60458250	시사IN News Magazi**	946,171	8	0
yonhaptwe**	147451838	연합**	332,637	24	0
Hongik**	481603067	홍익**	1,735	1	0
djuna**	151807455	dju**	83,020	37	0
woodyh**	118275399	Ha, Sungt**	9,638	7	0

누락된 사용자를 포함하여 총 172명의 트위터 사용자들의 속성 정보를 얻었다. 다음으로 사용자들의 Follwers_count, RT_count, Favorite_count를 이용하여 가중네트워크를 생성하기 위한 영향력 지수를 식 2-4를 통해 산출하였다. 사용자들의 영향력 지수는 표 3-13과 같고, 모든 사용자 172명의 영향력 지수는 <부록 A>에 수록하였다.

다음으로 사용자들이 쓴 텍스트 간의 유사도를 구하기 위해 Jaccard Index를 계산하여 주제 유사도를 구하였다. 표 3-14를 보면 사용자 아이디 '1117_06**'과 다른 사용자들 간의 주제 유사도를 구한 예시를 볼 수 있다. '1117_06**' 사용자와 '9px12**', 'Bad_systemerr**', 'Biael2**', 'ChaCha_**', 'Danteyeosan**', 'Glichel**', 'InHyun**', 'LiUriaISto**', 'Lians**', 'Lurence**', 'MaelVil**', 'Naloo10**', 'OUTIS_moer**', 'Peachsu**', 'PenKi**', 'RedKo**', 'SAY_**', 'Senoooo_**' 사용자는 주제 유사도가 1 값을 가진다. 이는 '1117_06**' 사용자와 위의 사용자들이 쓴 텍스트를 비교해보았을 때, 사용자들이 모두 같은 사용자를 리트윗했기 때문에 내용이 일치하기 때문으로 판단하였다. 주제 유사도 값이 0인 사용자들은 '1117_06**' 사용자와 전혀 다른 내용에 대해 언급하였기 때문으로 볼 수 있다. 이외에도 0과 1 사이의 값을 가지는 사용자들은 '1117_06**'과 비슷한 내용을 작성한 사용자들이라 할 수 있다.

표 3-13. '알파고와 이세돌의 바둑대국'에 해당하는 사용자들의 영향력 지수 일부

user_id	영향력 지수	user_id	영향력 지수
kbs01**	135.8084	youngeun02**	0.057104
sisain_edit**	50.16685	AndongYui_**	0.051462
HuffPostKor**	44.68213	Peachsu**	0.049827
woodyh**	28.74377	skcc_t**	0.049739
wikitr**	24.94205	nanpa_e**	0.049089
yonhaptwe**	18.02778	ktnbo**	0.048667
estim**	12.18829	zoltaen**	0.04582
PRESSIAN_ne**	7.963464	redpu**	0.04176
djuna**	5.646299	sangduck**	0.041338
leej**	3.886849	sekwoning**	0.040758
bstaeb**	1.653477	hgn3**	0.0406
iamkep**	1.32953	bluenote02**	0.039809
ooobbb**	1.107164	anne122**	0.039282
kscmylife**	0.846959	Arsenic**	0.037226
doo**	0.640479	minsu222**	0.035484
edaily_ne**	0.626875	hic**	0.034642
14O**	0.522001	PenKi**	0.034484
Xena_P**	0.420324	Niafilmuh**	0.034326
hellosamy**	0.360892	would_you_**	0.034062
cns53**	0.333395	Danteyeosan**	0.033904
yonhaptv_e**	0.225546	giantro**	0.030793
blueballoon0**	0.221297	dhzon**	0.028763
Hansarangn**	0.19546	casex90**	0.025362
aromayoungk**	0.174	Arken_K**	0.024044
JSY_worldmus**	0.172063	Biae12**	0.023569
Kyeongh**	0.167989	strongRengre**	0.019931
alleci**	0.165775	booknp**	0.017848
romnb**	0.162467	sokcu**	0.01773
hohoann**	0.153806	epfflzl**	0.017611
gkwqlsqud**	0.141098	ku993601**	0.016398
Hongik**	0.126191	tnwkor**	0.016293
Glichel**	0.099971	2nextdo**	0.015924
sangseek_k**	0.097967	OUTIS_moer**	0.015185
Lians**	0.077035	Venia_1**	0.015133
ellie**	0.0725	t_o**	0.014764
gksektha**	0.070472	lianfly02**	0.013604
maru51**	0.064064	tr62**	0.013393
rlarudf**	0.062726	blupe_**	0.012707
chogunwo**	0.058738	BaobabDunc**	0.012127

표 3-14. 트위터 사용자 '1117_06**'과 다른 사용자들의 주제 유사도 결과 중 일부

user_screen_name	user_screen_name	주제 유사도
1117_06**	119_112_1**	0
1117_06**	14O**	0.235294
1117_06**	2nextdo**	0
1117_06**	4444_**	0
1117_06**	9px12**	1
1117_06**	AndongYui_**	0
1117_06**	Arken_K**	0.1
1117_06**	Arsenic**	0
1117_06**	Bad_systemerr**	1
1117_06**	BaobabDunc**	0
1117_06**	Biae12**	1
1117_06**	ChaCha_**	1
1117_06**	CodeGenera**	0.083333
1117_06**	CrowKing6**	0
1117_06**	Danteyeosan**	1
1117_06**	Ellian_lem**	0
1117_06**	Glichel**	1
1117_06**	Hansarangn**	0
1117_06**	Hongik**	0
1117_06**	HuffPostKor**	0.052632
1117_06**	InHyun**	1
1117_06**	JSY_worldmus**	0.066667
1117_06**	Kyeongh**	0.037037
1117_06**	LABENO**	0
1117_06**	LiUriaISto**	1
1117_06**	Lians**	1
1117_06**	Lurence**	1
1117_06**	MaelVil**	1
1117_06**	Naloo10**	1
1117_06**	Niafilmuh**	0
1117_06**	OUTIS_moe**	1
1117_06**	PRESSIAN_ne**	0
1117_06**	Peachsu**	1
1117_06**	PenKi**	1
1117_06**	RazerKomac**	0.285714
1117_06**	RedKo**	1
1117_06**	SAY_**	1
1117_06**	Senoooo_**	1
1117_06**	Seopomaanpi**	0
1117_06**	TX_INSPIRATI**	0

3.2.4 KeyplayerRank를 이용한 키플레이어 탐지 결과

‘알파고와 이세돌의 바둑대국’이라는 주제로 트위터 사용자 172명에 대해 KpRank를 적용한 결과는 표 3-15와 같다. 모든 사용자 172명의 KpRank 결과는 <부록 B>에 수록하였다. 표 3-15는 상위 약 5%의 값을 가지는 사용자들의 KpRank 값을 나타내고 있다. KpRank가 가장 높은 값을 가지는 트위터 사용자 아이디는 ‘kbs01**’이며 ‘sisain_edit**’, ‘HuffPostKor**’, ‘woodyh**’, ‘wikitr**’, ‘yonhaptwe**’, ‘Glichel**’, ‘sangseek_k**’, ‘Lians**’, ‘Peachsu**’ 순으로 KpRank값이 높게 나왔다. 상위권에 속하는 트위터 사용자 중 ‘sisain_edit**’, ‘HuffPostKor**’, ‘wikitr**’, ‘yonhaptwe**’는 언론사의 트위터 계정이었다. 언론사 트위터 계정 이외에 다른 6명의 트위터 사용자는 일반인의 트위터 계정이었다.

표 3-15. ‘알파고와 이세돌의 바둑대국’의 KpRank 상위 5%

Rank	user_screen_name	KpRank 값
1	kbs01**	0.066794
2	sisain_edit**	0.023878
3	HuffPostKor**	0.022095
4	woodyh**	0.018393
5	wikitr**	0.016138
6	yonhaptwe**	0.010643
7	Glichel**	0.008486
8	sangseek_k**	0.008485
9	Lians**	0.008476
10	Peachsu**	0.008465
⋮	⋮	⋮

상위 10위권의 사용자들의 속성 정보를 표 3-16에 정리하여 KpRank 결과와 함께 비교해보았다. 일반적으로 언론사의 소셜 미디어 계정들은 무수히 많은 팔로워를 가지고 있다. 표 3-16을 보면 상위권에 있는 언론

사들은 다른 일반인들과 비교했을 때 상당히 높은 팔로워 수를 가진다. 따라서 KpRank를 구할 때 언론사들의 영향력 지수는 다른 일반인들보다 더 가중되었다. 그런데도 'kbs01**' 사용자가 1순위로 뽑힌 것은 리트윗된 횟수가 언론사들에 비해 월등히 많았기 때문이다. 7위부터 10위까지는 'kbs01**'의 텍스트를 리트윗한 사용자들이다. 해당 사용자들은 팔로워 수가 다른 사용자들에 비해 아주 높은 편도 아니고 리트윗 횟수도 0이지만, 가장 영향력 있는 'kbs01**'의 트윗을 리트윗했기 때문에 상위 5%에 든 것으로 판단하였다. KpRank가 페이지랭크 이론에 바탕을 두기 때문에 이러한 현상이 발생할 수 있다.

본 연구에서 탐지하고자 하는 키플레이어는 다른 사용자들에게 영향력이 강하고 주제에 특화된 사람을 찾는 것이었다. KpRank를 통해 구해진 1위의 사용자 'kbs01**'를 살펴보면 다른 사용자들이 'kbs01**'이 쓴 텍스트에 관심을 가지면서 리트윗을 많이 했기 때문에 소통이 활발히 됐다는 것을 알 수 있으며, 리트윗이 활발히 된 내용이 '알파고와 이세돌의 바둑대국'과 밀접한 내용임을 통해 이 주제에서의 키플레이어라고 할 수 있다.

표 3-16. 상위 5% 사용자들의 속성 정보

Rank	user_screen_name	followers_count	RT_count	Favorite_count
1	kbs01**	487	3,924	0
2	sisain_edit**	946,167	7.5	0
3	HuffPostKor**	821,282	29.33	0
4	woodyh**	9,637	816	0
5	wikitr**	418,292	80	0
6	yonhaptwe**	332,656	14	0
7	Glichel**	1,896	0	0
8	sangseek_k**	1,858	0	0
9	Lians**	1,461	0	0
10	Peachsu**	945	0	0

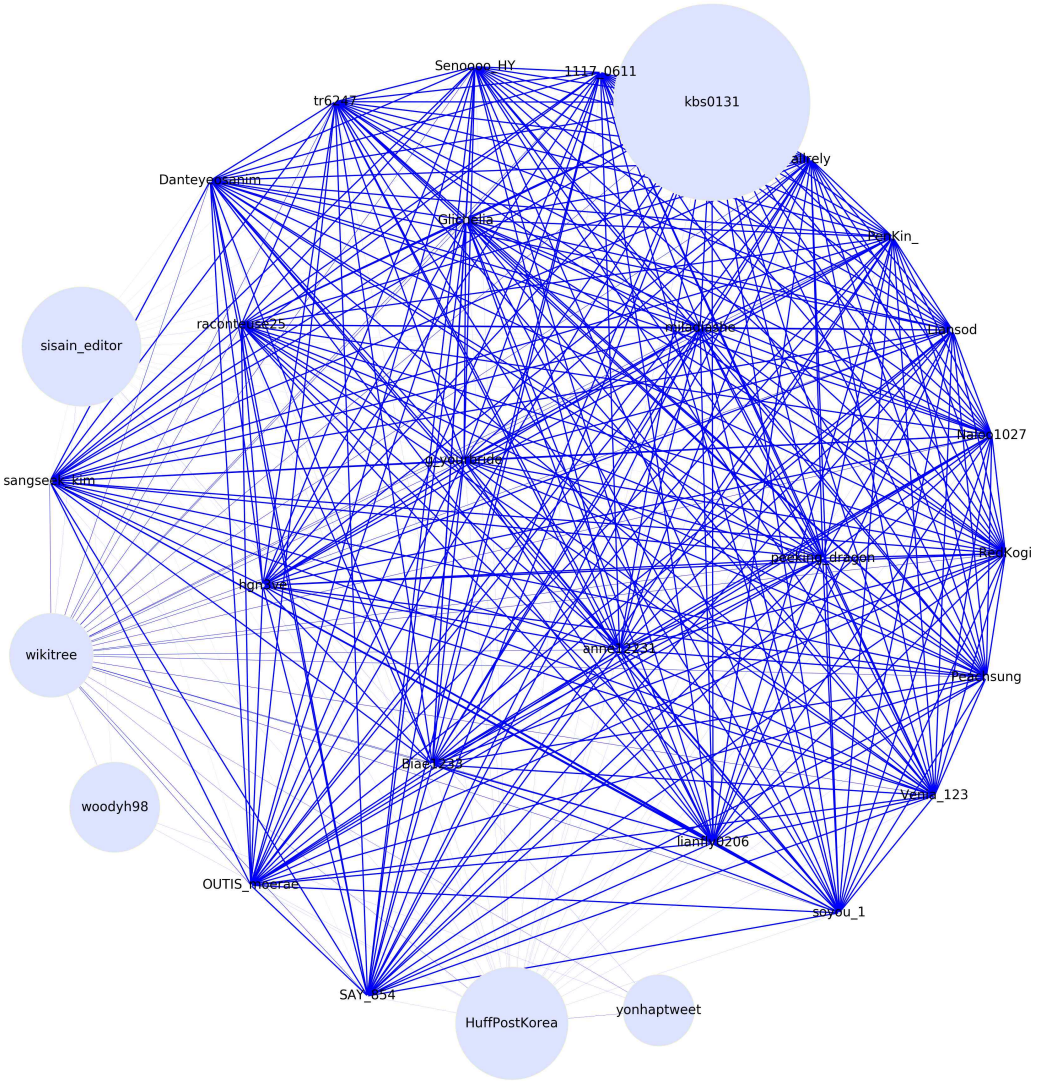

그림 3-4. ‘알파고와 이세돌의 바둑대국’에 해당하는 상위 30명 사용자의 KpRank 네트워크

그림 3-4는 상위 30명의 사용자를 네트워크로 표현한 모습이다. 이때 노드가 가장 큰 ‘kbs01**’이 키플레이어다. 여기서 링크 가중치는 주제 유사도이고, 노드 가중치는 Interaction Score이다. 링크 색이 진한색으로 보이는 값은 ‘1’로 네트워크 내에서 주제 유사도가 1인 링크가 유난히 많았다. 이러한 현상이 생긴 이유는 사용자들 간에 리트윗을 많이 하여 텍스트가 아예 일치하는 경우가 많이 생겼기 때문이다.

<부록 C>와 같이 영향력 지수와 주제 유사도 중 한 가지만을 고려하여 KpRank를 구하였을 때와 동시에 모두 고려하였을 때 사용자가 속하는 순위는 달라진다. 예를 들어, 'OUTIS_moer**'는 영향력 지수만을 고려했을 때 72위에 해당하는 사용자이다. 주제적 유사성을 고려하지 않고 영향력 측면에서만 봤을 때, 이 사용자는 모든 사용자 중에서 중간 정도의 영향력을 가졌다고 할 수 있다. 반대로 영향력 지수를 고려하지 않고 주제 유사도만 고려하여 KpRank를 구하였을 때, 이 사용자는 39위에 있다. '알파고와 이세돌의 바둑대국'이라는 주제에 관하여 많이 언급하므로 영향력 지수만을 고려한 순위보다는 높은 순위에 위치하는 것을 볼 수 있다. 그러나 <부록 B>와 같이 본 연구에서 제안하는 두 가지를 동시에 고려하여 KpRank를 구하였을 때, 'OUTIS_moer**'는 16위로 상위 10% 안에 드는 사용자였다. 이것은 '알파고와 이세돌의 바둑대국'이라는 주제 내에서 이 사용자가 가진 영향력과 다른 사용자들과의 주제적 유사성을 모두 고려한 결과로 한 가지만을 고려하였을 때보다 순위가 높다. 그 이유는 키플레이어인 'kbs01**'와 주제적 밀접한 관련성을 맺으면서 이 사용자가 주제에 미치는 상대적인 영향력이 높아졌기 때문이다.

이처럼 한 가지만을 반영하면 찾고자 하는 목적에 따라 편향된 결과를 얻을 수 있다. 하지만 동시에 고려하여 키플레이어를 찾고자 한다면 특정 주제에서 영향력 있는 사용자를 찾을 수 있다.

4. KeyplayerRank 활용 방안

본 연구에서 제안하는 KpRank는 다양한 측면에서 활용될 수 있다.

첫 번째, 트위터가 아닌 다른 소셜 미디어에도 적용될 수 있다. KpRank 식에 들어가는 두 가지 지수는 사용자의 속성 정보에 바탕을 두었기 때문에 다른 소셜 미디어에서 사용자의 정보를 스크래핑할 수 있다면 적용하기 용이하다. 먼저 주제 유사도는 사용자가 작성한 text에 기반하고, 영향력 지수는 리트윗과 favorite 그리고 팔로워 수를 반영하는데, 이것은 활용하고자 하는 소셜 미디어 사용자의 속성 정보를 통해 획득가능한 정보이다. 예를 들어 소셜 미디어 중 사람들이 많이 이용하는 인스타그램(Instagram)에 KpRank를 적용해보았다. 인스타그램은 그림 4-1, 4-2, 4-3과 같이 생겼다. 트위터와 같이 사용자의 아이디(user_screen_name)가 있고 그 아래에는 사용자의 별칭(user_id)이 있다. 그림 4-1의 사용자 아이디는 'minseonieyong'이고 사용자가 설정해놓은 별칭은 'M.Kim'이다. 또한, 트위터에서의 favorite은 인스타그램에서 '좋아요'와 유사하고, 팔로워 수도 사용자의 정보에 나와 있다(그림 4-1, 4-2). 리트윗 기능이 인스타그램에는 없지만 이와 유사한 기능인 '리그램(regram)' 기능이 있다(그림 4-3). '리그램' 기능은 트위터의 리트윗과 유사하게 상대방의 사진과 글을 자신의 인스타그램 계정에 다시 한번 게시하는 것을 의미한다. 따라서 사용자 정보만 알 수 있다면 KpRank식에 대입하여 인스타에서 주제별로 가장 영향력이 큰 키플레이어를 찾을 수 있다.

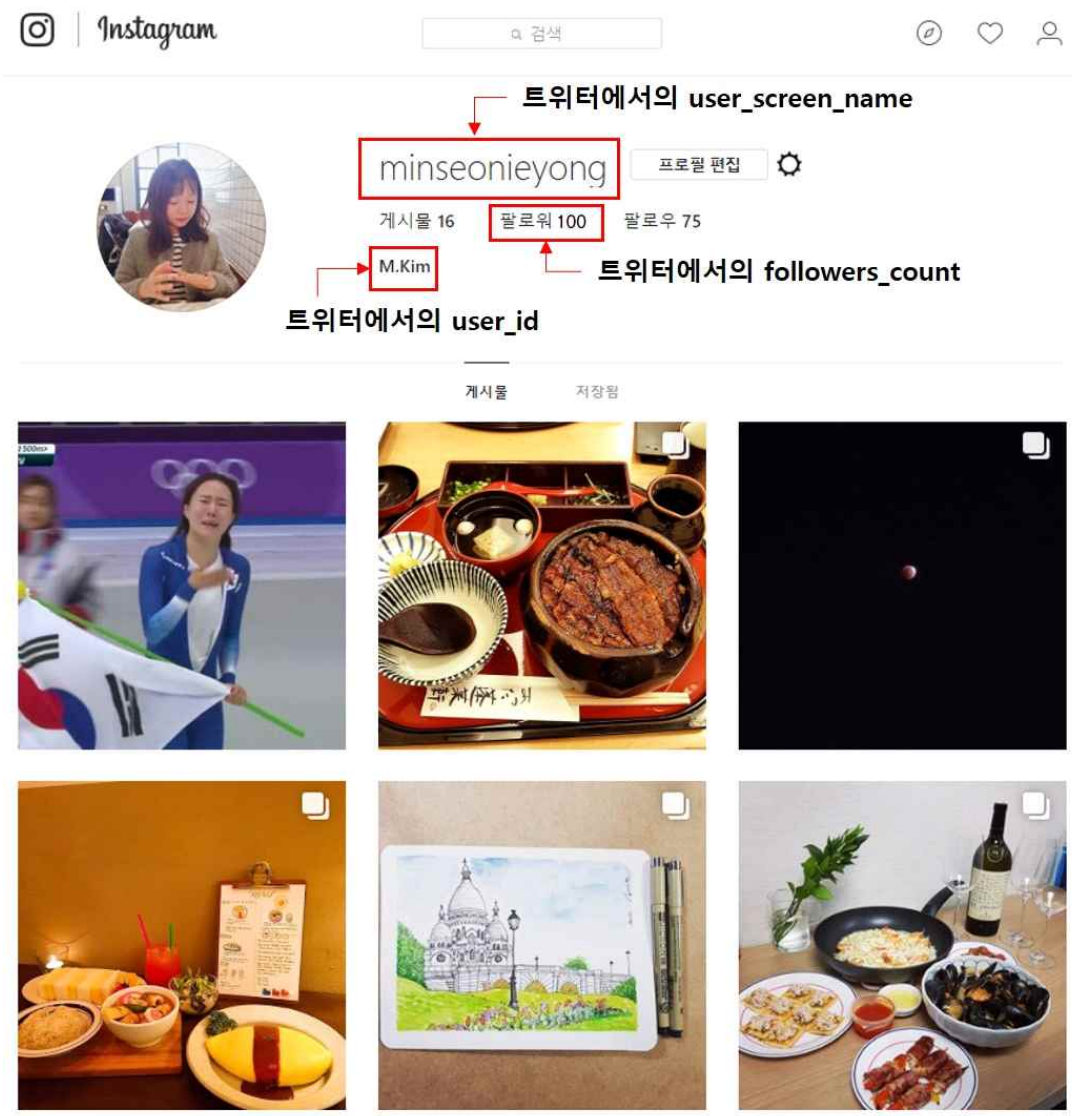

그림 4-1. 인스타그램의 사용자 정보(PC 화면)

그림 4-2. 인스타그램 텍스트의 정보(PC 화면)

그림 4-3. 인스타그램의 리그램 기능
(모바일 화면)

두 번째, 실시간으로 특정 공간에서의 키플레이어를 탐지하는 서비스에 적용할 수 있다. 본 연구에서는 트위터 사용자의 위치 정보를 고려하지 않았다. 하지만 향후 연구에 Streaming API로 트위터를 실시간으로 스크래핑하고 location 정보를 추가하면 실시간으로 올라오는 트위터들 사이의 관계를 분석하여 특정 주제와 공간에서의 키플레이어를 찾을 수 있다. 향후 제공할 수 있는 서비스 화면을 제시한다면 다음 그림 4-5와 같다.

마지막으로 특정 주제에서의 키플레이어를 찾는 것은 인플루언서 마케팅이나 정책 결정에 있어 활발하게 사용될 수 있다. 현재 LG전자의 경우 'THE BLOGer'라는 홈페이지를 운영하고 있다. 'THE BLOGer'는 유명한 파워블로거들을 모집하여 LG전자의 제품을 홍보하고 있다(그림 4-4). 이처럼 현재 기업들은 키플레이어와의 관계를 적극적으로 구축함으로 PR 및 마케팅 효과를 높이고 있다. 만약 KpRank를 이용한다면 기업에서 홍보하고자 하는 제품이 포함되는 주제에서 상위권의 순위를 가지는 사용자들과 접촉하여 효율적으로 인플루언서 마케팅 할 수 있다는 장점이 있다.

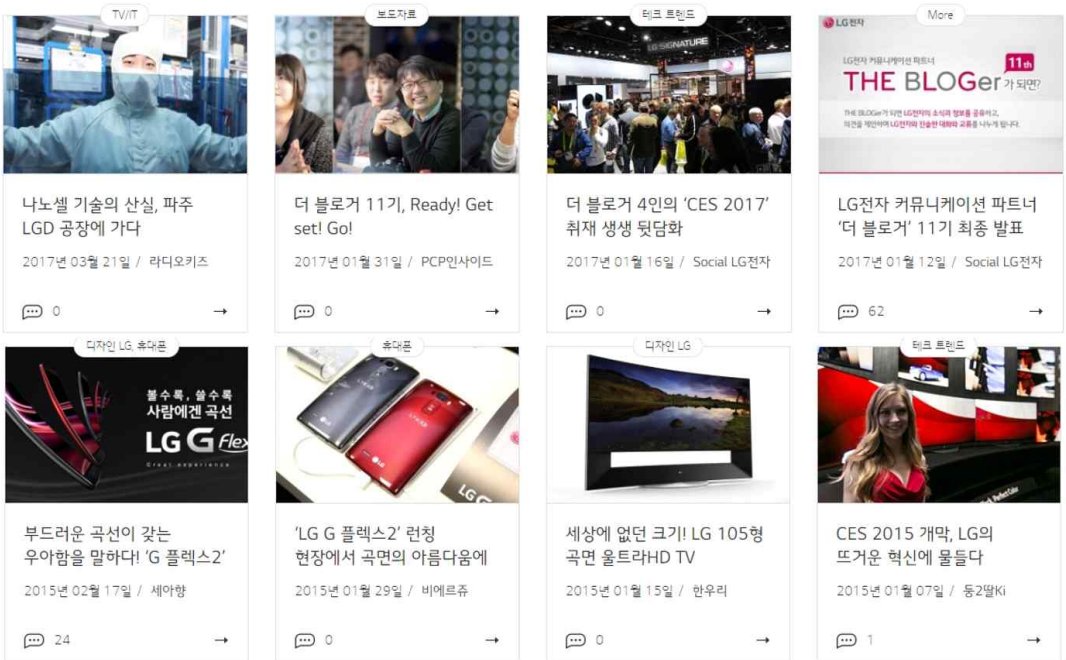

그림 4-4. LG전자의 'THE BLOGer' 화면 중 일부

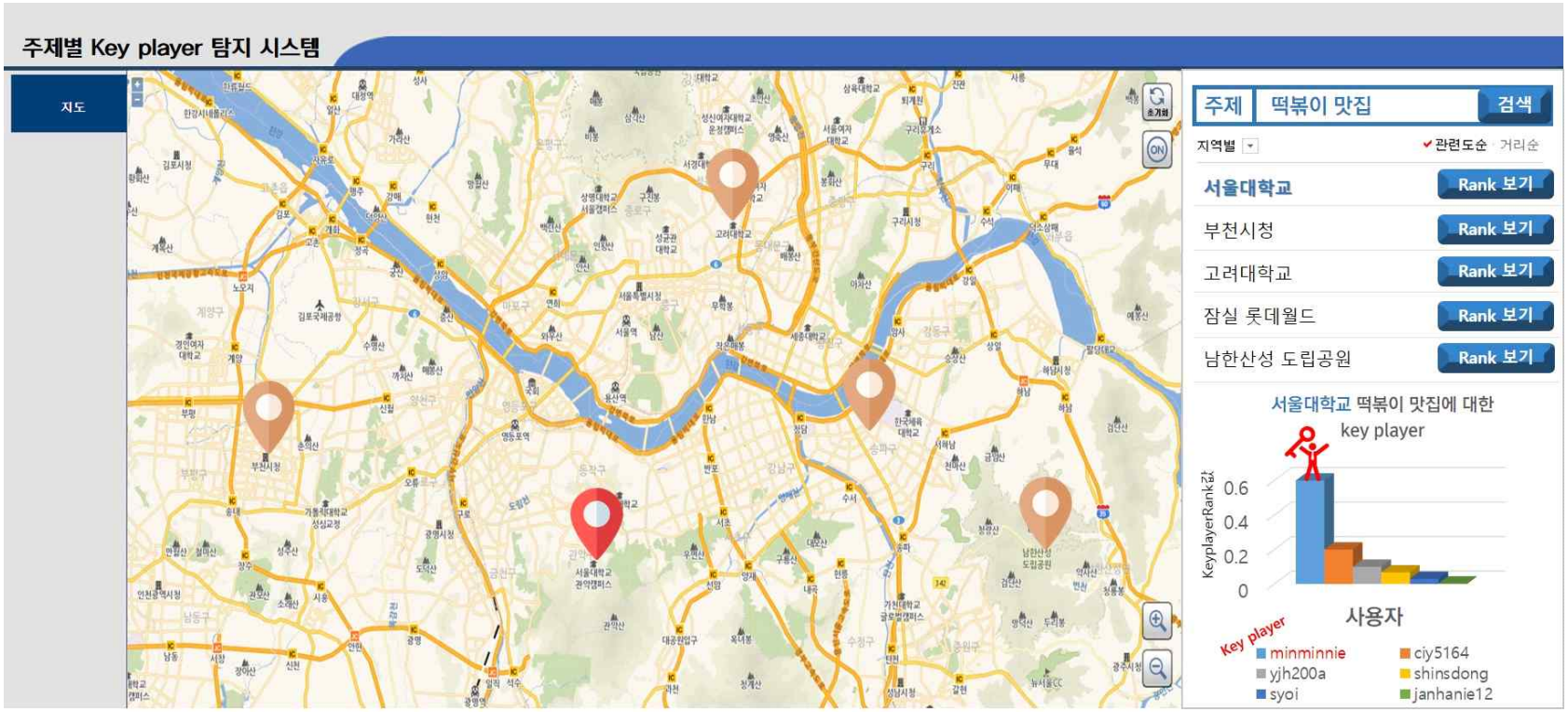

그림 4-5. 본 연구에서 제안하는 실시간 공간 및 주제별 키플레이어 탐지 서비스 화면 예시

5. 결론

인터넷과 스마트 기기를 활용하여 손쉽게 소셜 미디어를 접하면서 소셜 미디어 사용률은 매년 꾸준히 증가하고 있다. 소셜 미디어를 이용하는 다양한 목적 중에서 정보 습득과 소통이 가장 높은 비율을 차지한다. 소셜 미디어 사용자들이 필요한 정보를 얻는 데에 영향력 있는 사람의 의견이 상당한 부분을 차지하고 있다. 이처럼 키플레이어는 소셜 미디어 사용자들의 의사결정에 큰 영향을 주고 있다. 따라서 소셜 미디어상의 영향력자를 탐지하고자 하는 연구가 다양한 방법으로 이루어지고 있으며, 본 연구에서는 소셜 미디어상의 주제별 키플레이어를 탐지하는 KpRank 기법을 제안하였다. 제안한 기법은 KpRank에 영향력 지수와 주제 유사도를 동시에 고려하여 한 지표에 편향되지 않는 키플레이어를 탐지하였다. 이를 통해 두 지표를 단순 결합하여 키플레이어를 도출하였을 때 한 지표에 편향되는 문제점을 보완할 수 있다.

이에 본 연구는 영향력 지수와 주제 유사도를 적용한 KpRank를 제안한다. 영향력 지수를 통해 한 사용자가 다른 사용자들에게 끼치는 파급력, 영향력을 탐지하도록 하며, 주제 유사도를 통해 주제에 특화된 키플레이어를 탐지하도록 하였다. 그리고 주제별 키플레이어를 탐지하기 위해 토픽모델링 기법인 LDA 알고리즘을 사용하여 뉴스 기사에서 주제를 추출하고 SVM을 통해 주제와 관련성이 있는가를 판별하여 주제에 특화된 키플레이어를 탐지할 수 있도록 하였다.

약 1년간의 트위터 데이터를 이용하여 '알파고와 이세돌의 바둑대국'이라는 주제로 KpRank를 구한 결과 사용자 'kbs01**'이 키플레이어로 도출되었다. 다른 사용자들이 'kbs01**'이 쓴 텍스트를 많이 리트윗하는 것을 통해 이 사용자의 영향력을 알 수 있으며, 리트윗이 활발히 된 내용이 '알파고와 이세돌의 바둑대국'과 밀접한 내용임을 통해 이 주제에서의 키플레이어라고 판단하였다.

본 연구의 의의는 다음과 같다. 첫째, 특정 주제에서의 키플레이어를 탐지함으로써 그 주제에 관해 관심 있는 사람들이 수월하게 정보를 습득

하게 해줄 수 있다. 주제 유사도나 영향력 지수 둘 중 한 가지만을 고려하여 키플레이어를 탐지하는 기존 기법과는 달리, 본 연구에서 제안한 KpRank 기법을 통해 특정한 주제에 관심이 있고 그 주제에서의 영향력 있는 키플레이어를 찾을 수 있다. 따라서 본 연구는 소셜 미디어상에서 키플레이어를 찾고자 하는 목적에 따라 사용자가 선택할 수 있는 새로운 방법을 제안하였다.

둘째, 트위터만이 아닌 다른 소셜 미디어에도 적용 가능하다는 점에서 의의가 있다. 본 연구에서 제안한 KpRank는 소셜 미디어 사용자의 특성을 반영하였기 때문에 트위터 이외의 소셜 미디어에도 적용 가능하며 이를 통해 여러 소셜 미디어에서의 주제별 키플레이어도 탐지할 수 있다. 또한, 향후에는 실시간으로 제공되는 Streaming API와 공간적(location) 속성을 이용하여 해당 공간에서의 주제별 키플레이어를 찾는 서비스를 제공할 수 있다. 이외에도 기업에서 홍보하고자 하는 제품을 홍보하는 수단으로 키플레이어를 찾는데 효율적으로 활용될 수 있으며, 이는 인플루언서 마케팅에 효과적이라고 할 수 있다.

본 연구의 한계점으로는 첫 번째, 키워드에 해당하는 텍스트를 불러올 때 글의 맥락을 파악하지 않고 키워드에 매칭되는 텍스트를 불러오기 때문에 키워드가 존재하지 않거나 잘못 추출되는 텍스트들이 존재하였다. 이러한 문제는 키워드의 앞뒤 단어를 고려하고 텍스트 전체의 문맥도 고려하는 방법을 추가하여 개선해야 할 것이다. 두 번째, 주제 유사도를 구할 때 두 사용자가 같은 사용자를 리트윗한 경우 사용자 간의 텍스트가 똑같았기 때문에 해당 값이 1이 나오는 문제가 발생할 수 있다. 트위터 특성상 리트윗을 하는 경우가 많으므로, 리트윗인 텍스트의 경우 다른 가중치를 주는 등의 보완이 필요할 것이다.

향후 연구에서는 시공간적 개념을 포함하는 KpRank로 발전시키고자 하여, 주제별 키플레이어를 찾을 때 해당 주제가 가장 이슈가 되었던 시간이나 공간 등을 반영하면 조금 더 정확한 키플레이어를 탐지할 수 있을 것으로 생각한다.

참 고 문 헌

http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank (접속일 : 2018년 2월 14일)

http://www.blogher.com (접속일 : 2018년 5월 10일)

http://www.ktword.co.kr/abbr_view.php?m_temp1=4725 (접속일 : 2018년 5월 20일)

http://www.shineware.co.kr/products/komoran (접속일 : 2018년 5월 4일)

http://www.zdnet.com/article/social-networking-influence-followers-and-nexus-leaders/ (접속일 : 2018년 2월 10일)

https://simplymeasured.com/blog/how-to-define-identify-and-engage-social-media-influencers-for-your-brand/#sm.0000e47710pu0dtqynp11whpkqz5b (접속일 : 2018년 2월 1일)

https://www.slideshare.net/Altimeter/the-rise-of-digital-influence (접속일 : 2018년 5월 1일)

강애띠, 2016, 트윗에서 추출한 스트레스 감성과 토픽의 공간적 특성 연구, 이화여자대학교 대학원 박사학위 논문.

곽해운, 이창현, 박호성, 문수복, 2011, 트위터는 소셜 네트워크인가?-네트워크 구조와 정보 전파의 관점, 서울대학교 언론정보연구소, 제48권, 제1호, pp. 87-113.

김규하, 박철용, 2015, 토픽모형 및 사회연결망 분석을 이용한 한국데이터정보과학회지 영문초록 분석, 한국데이터 정보과학 학회지, 제26권 제1호, pp. 151-159.

김윤화, 2015, SNS(소셜네트워크서비스) 이용 추이 및 이용행태 분석, KISDI Stat Report, 제15권, 제3호, pp. 7-12.

김정호, 김명규, 차명훈, 인주호, 채수환, 2010, 선택적 자질 차원 축소를 이용한 최적의 지도적 LSA 방법, 감성과학, 제13권, 제1호, pp. 47-60.

김하진, 정효정, 송민, 2014, 토픽모델링을 통한 저자명 식별 성능 비교, 한국정보관리학회 학술대회 논문집, pp. 149-152.

김형석, 박민석, 강필성, 2016, 토픽모델링과 사회연결망을 통한 딥러닝 연구동향 분석, 한국경영과학회 학술대회논문집, pp. 1877-1899.

남춘호, 2016, 일기자료 연구에서 토픽모델링 기법의 활용 가능성 검토, 서울대학교 비교문화연구소, 제22권, 제1호, pp. 89-135.

박자현, 송민, 2013, 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석, 정보관리학회지, 제30권 제1호, pp. 7-32.

박호성, 곽해운, 차미영, 문수복, 2010, 소셜 네트워크에서의 인플루엔셜 랭킹, 정보과학회지, 제28권, 제3호, pp. 24-30.

배정환, 손지은, 송민, 2013, 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석, 지능정보연구, 제19권, 제3호, pp. 141-156.

이원태, 정부연, 2011, 소셜플랫폼의 사회적 영향력 분석 및 발전방향 연구, 정책연구, 정보통신정책연구원.

정영미, 임혜영, 2000, SVM 분류기를 이용한 문서 범주화 연구, 정보관리학회지, 제17권, 제4호, pp. 229-248.

진설아, 허고은, 정유경, 송민, 2013, 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구, 정보관리학회지, 제30권, 제1호, pp. 285-302.

차윤정, 이지혜, 최지은, 김희웅, 2015, 소셜 미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원, 지식경영연구, 제16권, 제4호, pp. 69-87.

한국인터넷진흥원, 2013, 2013년 모바일인터넷이용실태조사, pp. 1-306.

Blei, D. M., 2012, Probabilistic topic models, *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84.

Blei, David M., Andrew Ng, Michael Jordan, 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Brin, S., Page, L., 1998, The anatomy of a large-scale hypertextual web search engine, *Computer networks and ISDN systems*, Vol. 30, pp. 107-117.

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P. K., 2010, Measuring user influence in twitter: The million follower fallacy, *ICWSM*, Vol. 30, No. 10, pp. 10-17.

Cortes, C., Vapnik, V., 1995, Support-vector networks, *Machine learning*, Vol. 20, No. 3, pp. 273-297.

Dumais, S., Platt, J., Heckerman, D., Sahami, M., 1998, Inductive learning algorithms and representations for text categorization, *In Proceedings of the seventh international conference on Information and knowledge management*, ACM, pp. 148-155.

Ghosh, D., Guha, R., 2013, What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System, *Cartography and geographic information science*, Vol. 40, No. 2, pp. 90-102.

Java, A., Song, X., Finin, T., Tseng, B., 2007, Why we twitter: understanding microblogging usage and communities, *In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis,* ACM, pp. 56-65.

Joachims, T., 1998, Text categorization with support vector machines: Learning with many relevant featuresm, I*n European conference on machine learning*, pp. 137-142.

Katz, E., Lazarsfeld, P. F., 1966, Personal Influence, The part played by people in the flow of mass communications, *Transaction Publishers*.

Kleinberg, J. M., 1998, Authoritative sources in a hyperlinked environment, *In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

Kwak, H., Lee, C., Park, H., Moon, S., 2010, What is Twitter, a social network or a news media?, *In Proceedings of the 19th international conference on World wide web,* ACM, pp. 591-600.

Leavitt, A., Burchard, E., Fisher, D., Gilbert, S., 2009, The influentials: New approaches for analyzing influence on twitter, *Web Ecology Project*, Vol. 4, No. 2, pp. 1-18.

Livne, A., Simmons, M. P., Adar, E., Adamic, L. A., 2011, The Party Is Over Here: Structure and Content in the 2010 Election, *ICWSM*, Vol. 11, pp. 17-21.

Malcom Gladwell, 2000, The Tipping Point: How Little Things Can Make a Big Difference, Little, Brown.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001, Birds of a feather: Homophily in social networks, *Annual review of sociology*, Vol. 27, No. 1, pp. 415-444.

Real, R., Vargas, J. M.., 1996, The probabilistic basis of Jaccard's index of similarity, *Systematic biology*, Vol. 45, No. 3, pp. 380-385.

Romero, R., Iglesias, E. L., Borrajo, L., 2015, A linear-RBF multikernel SVM to classify big text corpora, *BioMed research international*.

Scholkopf, B., Sung, K. K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997, Comparing support vector machines with

Gaussian kernels to radial basis function classifiers, *IEEE transactions on Signal Processing*, Vol. 45, No. 11, pp. 2758-2765.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Ungar, L. H., 2013, Characterizing Geographic Variation in Well-Being Using Tweets, *In ICWSM*, pp. 583-591.

Sriurai, W., 2011, Improving text categorization by using a topic model, *Advanced Computing*, Vol. 2, No. 6, pp. 21-27.

Sung, J., Moon, S., Lee, J. G., 2013, The influence in twitter: Are they really influenced?, *In BSI@ PAKDD/BSIC@ IJCAI*, pp. 95-105.

Weng, J., Lim, E. P., Jiang, J., He, Q., 2010, Twitterrank: finding topic-sensitive influential twitterers, *In Proceedings of the third ACM international conference on Web search and data mining*, pp. 261-270.

부 록

<부록 A> ‘알파고와 이세돌의 바둑대국’에 해당하는
사용자들의 영향력 지수

<부록 B> ‘알파고와 이세돌의 바둑대국’의
KpRank 결과

<부록 C> 영향력 지수 또는 주제 유사도만을 고려한
‘알파고와 이세돌의 바둑대국’의 KpRank 결과

<부록 A> '알파고와 이세돌의 바둑대국'에 해당하는 사용자들의 영향력 지수

표 A-1. '알파고와 이세돌의 바둑대국'에 해당하는 사용자들의 영향력 지수

user_id	영향력 지수	user_id	영향력 지수
kbs01**	135.8084	maru51**	0.064064
sisain_edit**	50.16685	rlarudf**	0.062726
HuffPostKor**	44.68213	chogunwo**	0.058738
woodyh**	28.74377	youngeun02**	0.057104
wikitr**	24.94205	AndongYui_**	0.051462
yonhaptwe**	18.02778	Peachsu**	0.049827
estim**	12.18829	skcc_t**	0.049739
PRESSIAN_ne**	7.963464	nanpa_e**	0.049089
djuna**	5.646299	ktnbo**	0.048667
leej**	3.886849	zoltaen**	0.04582
bstaeb**	1.653477	redpu**	0.04176
iamkep**	1.32953	sangduck**	0.041338
ooobbb**	1.107164	sekwoning**	0.040758
kscmylife**	0.846959	hgn3**	0.0406
doo**	0.640479	bluenote02**	0.039809
edaily_ne**	0.626875	anne122**	0.039282
14O**	0.522001	Arsenic**	0.037226
Xena_P**	0.420324	minsu222**	0.035484
hellosamy**	0.360892	hic**	0.034642
cns53**	0.333395	PenKi**	0.034484
yonhaptv_e**	0.225546	Niafilmuh**	0.034326
blueballoon0**	0.221297	would_you_**	0.034062
Hansarangn**	0.19546	Danteyeosan**	0.033904
aromayoungk**	0.174	giantro**	0.030793
JSY_worldmus**	0.172063	dhzon**	0.028763
Kyeongh**	0.167989	casex90**	0.025362
alleci**	0.165775	Arken_K**	0.024044
romnb**	0.162467	Biae12**	0.023569
hohoann**	0.153806	strongRengre**	0.019931
gkwqlsqud**	0.141098	booknp**	0.017848
Hongik**	0.126191	sokcu**	0.01773
Glichel**	0.099971	epfflzl**	0.017611
sangseek_k**	0.097967	ku993601**	0.016398
Lians**	0.077035	tnwkor**	0.016293
ellie**	0.0725	2nextdo**	0.015924
gksektha**	0.070472	OUTIS_moer**	0.015185

user_id	영향력 지수	user_id	영향력 지수
Venia_1**	0.015133	hhhoney_**	0.004693
t_o**	0.014764	yuha_07**	0.004693
lianfly02**	0.013604	tonytony05**	0.00464
tr62**	0.013393	sang_uk_g**	0.004535
blupe_**	0.012707	9px12**	0.004482
BaobabDunc**	0.012127	naksan**	0.004429
dorami_20**	0.011758	danmu_**	0.004324
Naloo10**	0.011705	snowhoo**	0.004271
Senoooo_**	0.010915	MaelVil**	0.004218
RedKo**	0.010862	sunkitty**	0.00406
miladias**	0.010335	clacladr**	0.003796
ikkiS2sh**	0.010282	nunlt**	0.003796
kuin01**	0.010176	CrowKing6**	0.003744
dambae_j**	0.009913	_14927**	0.003691
g_yourbri**	0.009333	runaka**	0.003691
ruine**	0.008964	lovefu**	0.003585
allre**	0.008542	pippyap8**	0.003427
rainpop**	0.008489	laina9610**	0.003269
soyou**	0.008436	aeghilmnopru**	0.003216
TX_INSPIRATI**	0.008331	buncoch**	0.003164
1117_06**	0.007435	LiUriaISto**	0.003058
weevil1**	0.007382	Yuko_A**	0.003058
CodeGenerat**	0.007329	re_gul_n**	0.003058
kiR_ok**	0.00696	ksy445**	0.002945
pjhmyuko**	0.00696	Lurence**	0.0029
Ellian_lem**	0.006538	camilla05**	0.002795
SAY_8**	0.006433	seasoning_sa**	0.002795
peeking_drag**	0.006327	kokakan**	0.002689
raconteuse**	0.006222	InHyun**	0.002531
ramrump12**	0.006169	directorz**	0.002425
bhASAnti**	0.005958	TheCatcherNe**	0.002373
swap_ye**	0.005958	bu16**	0.002267
aaswf**	0.005905	hoopbugintsul**	0.002109
Seopomaanpi**	0.005536	selll1**	0.002109
Bad_systemerr**	0.005325	neet_seek**	0.001951
sadarl**	0.00522	ChaCha_**	0.001793
yoonjuum**	0.00522	ogwai9**	0.00174
whuiyin**	0.005062	LABENO**	0.001529
naengmyeonki**	0.005009	whin3**	0.001529
peTales**	0.004956	119_112_1**	0.001371
RazerKomac**	0.004693	soulmateheali**	0.001213

user_id	영향력 지수
luka_10**	0.00116
raccoon_clo**	0.001107
taehyeon1**	0.001055
sonju**	0.000949
Ungo**	0.000844
chaeyoung_**	0.000738
songhee_**	0.000738
k_kanm**	0.000685
4444_**	0.000633
trans_gam**	0.00058
tigerblack5**	0.000422
sweetne2**	0.000369
yummy_star_s**	0.000369
nothingboo**	0.000316
fbtpfld**	0.000264
poca_preghie**	0.000264
imnotcompl**	0.000158
cosse01**	0.000053

<부록 B> '알파고와 이세돌의 바둑대국'의 KpRank 결과

표 B-1. '알파고와 이세돌의 바둑대국'의 KpRank 결과

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
1	kbs01**	0.066794	37	tonytony05**	0.008445
2	sisain_edit**	0.023878	38	9px12**	0.008445
3	HuffPostKor**	0.022095	39	naksan**	0.008445
4	woodyh**	0.018393	40	danmu_**	0.008445
5	wikitr**	0.016138	41	MaelVil**	0.008445
6	yonhaptwe**	0.010643	42	nunlt**	0.008445
7	Glichel**	0.008486	43	runaka**	0.008445
8	sangseek_k**	0.008485	44	pippyap8**	0.008445
9	Lians**	0.008476	45	aeghilmnopru**	0.008445
10	Peachsu**	0.008465	46	LiUriaISto**	0.008444
11	hgn3**	0.008461	47	Lurence**	0.008444
12	anne122**	0.00846	48	camilla05**	0.008444
13	PenKi**	0.008458	49	seasoning_sa**	0.008444
14	Danteyeosan**	0.008458	50	InHyun**	0.008444
15	Biae12**	0.008453	51	bu16**	0.008444
16	OUTIS_moer**	0.00845	52	hoopbugintsul**	0.008444
17	Venia_1**	0.00845	53	neet_seek**	0.008444
18	lianfly02**	0.008449	54	ChaCha_**	0.008444
19	tr62**	0.008449	55	whin3**	0.008444
20	Naloo10**	0.008448	56	luka_10**	0.008444
21	Senoooo_**	0.008448	57	Ungo**	0.008443
22	RedKo**	0.008448	58	trans_gam**	0.008443
23	miladias**	0.008448	59	sweetne2**	0.008443
24	g_yourbri**	0.008447	60	yummy_star_**	0.008443
25	allre**	0.008447	61	redpu**	0.006216
26	soyou**	0.008447	62	sangduck**	0.006216
27	1117_06**	0.008446	63	hic**	0.006213
28	SAY_8**	0.008446	64	epfflzl**	0.006206
29	peeking_drag**	0.008446	65	blupe_**	0.006204
30	raconteuse**	0.008446	66	dorami_20**	0.006204
31	ramrump12**	0.008446	67	ikkiS2sh**	0.006203
32	swap_ye**	0.008446	68	weevil1**	0.006202
33	Bad_systemer**	0.008445	69	kiR_ok**	0.006202
34	yoonjuum**	0.008445	70	Ellian_lem**	0.006201
35	hhhoney_**	0.008445	71	sang_uk_g**	0.0062
36	yuha_07**	0.008445	72	snowhoo**	0.0062

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
73	clacladr**	0.0062	114	bluenote02**	0.003161
74	_14927**	0.0062	115	would_you_**	0.003159
75	Yuko_A**	0.0062	116	giantro**	0.003157
76	kokakan**	0.0062	117	bhASAnti**	0.003147
77	directorz**	0.0062	118	aromayoungk**	0.003113
78	ogwai9**	0.006199	119	romnb**	0.003108
79	119_112_1**	0.006199	120	Hongik**	0.003033
80	raccoon_clo**	0.006199	121	ku993601**	0.003019
81	songhee_**	0.006199	122	casex90**	0.00289
82	k_kanm**	0.006199	123	tigerblack5**	0.002886
83	4444_**	0.006199	124	BaobabDunc**	0.002883
84	nothingboo**	0.006199	125	CrowKing6**	0.00257
85	estim**	0.005926	126	lovefu**	0.002466
86	PRESSIAN_ne**	0.005709	127	ktnbo**	0.002405
87	zoltaen**	0.005595	128	chogunwo**	0.002404
88	sekwoning**	0.005592	129	edaily_ne**	0.002387
89	RazerKomac**	0.005577	130	bstaeb**	0.002339
90	selll1**	0.005576	131	imnotcompl**	0.00228
91	djuna**	0.005447	132	TheCatcherNe**	0.002228
92	Xena_P**	0.004166	133	ooobbb**	0.002124
93	14O**	0.004086	134	doo**	0.002086
94	AndongYui_**	0.004013	135	soulmateheali**	0.00208
95	Niafilmuh**	0.004005	136	youngeun02**	0.002059
96	t_o**	0.003997	137	Arken_K**	0.002041
97	kuin01**	0.003995	138	peTales**	0.002039
98	ruine**	0.003995	139	fbtpfld**	0.002037
99	rainpop**	0.003995	140	buncoch**	0.002033
100	whuiyin**	0.003993	141	hohoann**	0.00203
101	re_gul_n**	0.003992	142	gkwqlsqud**	0.002025
102	LABENO**	0.003992	143	sokcu**	0.002021
103	chaeyoung_**	0.003991	144	gksektha**	0.001997
104	2nextdo**	0.003828	145	pjhmyuko**	0.001972
105	yonhaptv_e**	0.003647	146	ksy445**	0.00184
106	nanpa_e**	0.003606	147	cns53**	0.001822
107	aaswf**	0.003588	148	strongRengre**	0.00182
108	naengmyeonki**	0.003588	149	maru51**	0.001665
109	taehyeon1**	0.003586	150	Hansarangn**	0.001648
110	sonju**	0.003555	151	skcc_t**	0.001647
111	leej**	0.003389	152	tnwkor**	0.00163
112	kscmylife**	0.003326	153	Seopomaanp**	0.001523
113	sadarl**	0.003257	154	iamkep**	0.001485

Rank	user_screen_name	KpRank 값
155	Arsenic**	0.001427
156	booknp**	0.001406
157	cosse01**	0.001371
158	blueballoon0**	0.001346
159	Kyeongh**	0.001324
160	sunkitty**	0.001304
161	alleci**	0.001298
162	ellie**	0.001168
163	dambae_j**	0.001146
164	poca_preghie**	0.000998
165	TX_INSPIRATI**	0.000984
166	JSY_worldmu**	0.000922
167	CodeGenerat**	0.000911
168	dhzon**	0.000907
169	rlarudf**	0.000862
170	minsu222**	0.000833
171	laina9610**	0.000747
172	hellosamy**	0.000723

<부록 C> 영향력 지수 또는 주제 유사도만을 고려한 '알파고와 이세돌의 바둑대국'의 KpRank 결과

표 C-1. 영향력 지수만을 고려한 '알파고와 이세돌의 바둑대국'의 KpRank 결과

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
1	kbs01**	0.063872	37	maru51**	0.004974
2	sisain_edit**	0.026713	38	rlarudf**	0.004973
3	HuffPostKor**	0.024333	39	chogunwo**	0.004972
4	woodyh**	0.017418	40	youngeun02**	0.004971
5	wikitr**	0.015768	41	AndongYui_**	0.004969
6	yonhaptwe**	0.012768	42	Peachsu**	0.004968
7	estim**	0.010235	43	skcc_t**	0.004968
8	PRESSIAN_ne**	0.008401	44	nanpa_e**	0.004967
9	djuna**	0.007396	45	ktnbo**	0.004967
10	leej**	0.006633	46	zoltaen**	0.004966
11	bstaeb**	0.005664	47	redpu**	0.004964
12	iamkep**	0.005523	48	sangduck**	0.004964
13	ooobbb**	0.005427	49	sekwoning**	0.004964
14	kscmylife**	0.005314	50	hgn3**	0.004964
15	doo**	0.005224	51	bluenote02**	0.004963
16	edaily_ne**	0.005218	52	anne122**	0.004963
17	14O**	0.005173	53	Arsenic**	0.004962
18	Xena_P**	0.005129	54	minsu222**	0.004962
19	hellosamy**	0.005103	55	hic**	0.004961
20	cns53**	0.005091	56	PenKi**	0.004961
21	yonhaptv_e**	0.005044	57	Niafilmuh**	0.004961
22	blueballoon0**	0.005042	58	would_you_**	0.004961
23	Hansarangn**	0.005031	59	Danteyeosan**	0.004961
24	aromayoungk**	0.005022	60	giantro**	0.00496
25	JSY_worldmus**	0.005021	61	dhzon**	0.004959
26	Kyeongh**	0.005019	62	casex90**	0.004957
27	alleci**	0.005018	63	Arken_K**	0.004957
28	romnb**	0.005017	64	Biae12**	0.004956
29	hohoann**	0.005013	65	strongRengre**	0.004955
30	gkwqlsqud**	0.005007	66	booknp**	0.004954
31	Hongik**	0.005001	67	sokcu**	0.004954
32	Glichel**	0.00499	68	epfflzl**	0.004954
33	sangseek_k**	0.004989	69	ku993601**	0.004953
34	Lians**	0.00498	70	tnwkor**	0.004953
35	ellie**	0.004978	71	2nextdo**	0.004953
36	gksektha**	0.004977	72	OUTIS_moer**	0.004953

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
73	Venia_1**	0.004953	114	hhhoney_**	0.004948
74	t_o**	0.004953	115	yuha_07**	0.004948
75	lianfly02**	0.004952	116	tonytony05**	0.004948
76	tr62**	0.004952	117	sang_uk_g**	0.004948
77	blupe_**	0.004952	118	9px12**	0.004948
78	BaobabDunc**	0.004951	119	naksan**	0.004948
79	dorami_20**	0.004951	120	danmu_**	0.004948
80	Naloo10**	0.004951	121	snowhoo**	0.004948
81	Senoooo_**	0.004951	122	MaelVil**	0.004948
82	RedKo**	0.004951	123	sunkitty**	0.004948
83	miladias**	0.004951	124	nunlt**	0.004948
84	ikkiS2sh**	0.004951	125	clacladr**	0.004948
85	kuin01**	0.004951	126	CrowKing6**	0.004948
86	dambae_j**	0.00495	127	_14927**	0.004948
87	g_yourbri**	0.00495	128	runaka**	0.004948
88	ruine**	0.00495	129	lovefu**	0.004948
89	allre**	0.00495	130	pippyap8**	0.004948
90	rainpop**	0.00495	131	laina9610**	0.004948
91	soyou_1**	0.00495	132	aeghilmnopru**	0.004948
92	TX_INSPIRATI**	0.00495	133	buncoch**	0.004948
93	1117_06**	0.004949	134	LiUriaISto**	0.004948
94	weevil1**	0.004949	135	Yuko_A**	0.004948
95	CodeGenerat**	0.004949	136	re_gul_n**	0.004948
96	kiR_ok**	0.004949	137	ksy445**	0.004947
97	pjhmyuko**	0.004949	138	Lurence**	0.004947
98	Ellian_lem**	0.004949	139	camilla05**	0.004947
99	SAY_8**	0.004949	140	seasoning_sa**	0.004947
100	peeking_drag**	0.004949	141	kokakan**	0.004947
101	raconteuse**	0.004949	142	InHyun**	0.004947
102	ramrump12**	0.004949	143	directorz**	0.004947
103	bhASAnti**	0.004949	144	TheCatcherNe**	0.004947
104	swap_ye**	0.004949	145	bu16**	0.004947
105	aaswf**	0.004949	146	hoopbugintsul**	0.004947
106	Seopomaanpi**	0.004949	147	selll1**	0.004947
107	Bad_systemerr**	0.004948	148	neet_seek**	0.004947
108	sadarl**	0.004948	149	ChaCha_**	0.004947
109	yoonjuum**	0.004948	150	ogwai9**	0.004947
110	whuiyin**	0.004948	151	LABENO**	0.004947
111	naengmyeonki**	0.004948	152	whin3**	0.004947
112	peTales**	0.004948	153	119_112_1**	0.004947
113	RazerKomac**	0.004948	154	soulmateheali**	0.004947

Rank	user_screen_name	KpRank 값
155	luka_10**	0.004947
156	raccoon_clo**	0.004947
157	taehyeon1**	0.004947
158	sonju**	0.004947
159	Ungo**	0.004947
160	chaeyoung_**	0.004946
161	songhee_**	0.004946
162	k_kanm**	0.004946
163	4444_**	0.004946
164	trans_gam**	0.004946
165	tigerblack5**	0.004946
166	yummy_star_s**	0.004946
167	sweetne2**	0.004946
168	nothingboo**	0.004946
169	fbtpfld**	0.004946
170	poca_preghie**	0.004946
171	imnotcompl**	0.004946
172	cosse01**	0.004946

표 C-2. 주제 유사도만을 고려한 '알파고와 이세돌의 바둑대국'의 KpRank 결과

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
1	119_112_1**	0.008724	41	PenKi**	0.008052
2	4444_**	0.008724	42	RedKo**	0.008052
3	Ellian_lem**	0.008724	43	SAY_8**	0.008052
4	Yuko_A**	0.008724	44	Senoooo_**	0.008052
5	_14927**	0.008724	45	Ungo**	0.008052
6	blupe_**	0.008724	46	Venia_1**	0.008052
7	clacladr**	0.008724	47	aeghilmnopru**	0.008052
8	directorz**	0.008724	48	allre**	0.008052
9	dorami_20**	0.008724	49	anne122**	0.008052
10	epfflzl**	0.008724	50	bu16**	0.008052
11	hic**	0.008724	51	camilla05**	0.008052
12	ikkiS2sh**	0.008724	52	danmu_**	0.008052
13	k_kanm**	0.008724	53	g_yourbri**	0.008052
14	kiR_ok**	0.008724	54	hgn3**	0.008052
15	kokakan**	0.008724	55	hhhoney_**	0.008052
16	nothingboo**	0.008724	56	hoopbugintsul**	0.008052
17	ogwai9**	0.008724	57	kbs01**	0.008052
18	raccoon_clo**	0.008724	58	lianfly02**	0.008052
19	redpu**	0.008724	59	luka_10**	0.008052
20	sang_uk_g**	0.008724	60	miladias**	0.008052
21	sangduck**	0.008724	61	naksan**	0.008052
22	snowhoo**	0.008724	62	neet_seek**	0.008052
23	songhee_**	0.008724	63	nunlt**	0.008052
24	weevil1**	0.008724	64	peeking_drag**	0.008052
25	woodyh**	0.008724	65	pippyap8**	0.008052
26	runaka**	0.008052	66	raconteuse**	0.008052
27	9px12**	0.008052	67	ramrump12**	0.008052
28	Bad_systemerr**	0.008052	68	sangseek_k**	0.008052
29	Biae12**	0.008052	69	seasoning_sa**	0.008052
30	ChaCha_**	0.008052	70	soyou**	0.008052
31	Danteyeosan**	0.008052	71	swap_ye**	0.008052
32	Glichel**	0.008052	72	sweetne2**	0.008052
33	InHyun**	0.008052	73	tonytony05**	0.008052
34	LiUriaISto**	0.008052	74	tr62**	0.008052
35	Lians**	0.008052	75	trans_gam**	0.008052
36	Lurence**	0.008052	76	whin3**	0.008052
37	MaelVil**	0.008052	77	yoonjuum**	0.008052
38	Naloo10**	0.008052	78	yuha_07**	0.008052
39	OUTIS_moer**	0.008052	79	yummy_star_s**	0.008052
40	Peachsu**	0.008052	80	1117_06**	0.008052

Rank	user_screen_name	KpRank 값	Rank	user_screen_name	KpRank 값
81	RazerKomac**	0.006137	122	PRESSIAN_ne**	0.003401
82	sekwoning**	0.006137	123	peTales**	0.003334
83	selll1**	0.006137	124	youngeun02**	0.003334
84	wikitr**	0.006137	125	fbtpfld**	0.003334
85	zoltaen**	0.006137	126	imnotcompl**	0.003331
86	AndongYui_**	0.006069	127	gkwqlsqud**	0.003298
87	LABENO**	0.006069	128	hohoann**	0.003298
88	Niafilmuh**	0.006069	129	chogunwo**	0.003263
89	Xena_P**	0.006069	130	pjhmyuko**	0.003199
90	chaeyoung_**	0.006069	131	gksektha**	0.003199
91	kuin01**	0.006069	132	edaily_ne**	0.003187
92	rainpop**	0.006069	133	TheCatcherNe**	0.003173
93	ruine**	0.006069	134	sisain_edit**	0.00315
94	re_gul_n**	0.006069	135	tigerblack5**	0.003142
95	t_o**	0.006069	136	ku993601**	0.003015
96	whuiyin**	0.006069	137	leej**	0.002919
97	2nextdo**	0.005576	138	soulmateheali**	0.002886
98	aaswf**	0.005079	139	doo**	0.002868
99	naengmyeonki**	0.005079	140	sokcu**	0.002746
100	nanpa_e**	0.005079	141	strongRengre**	0.002738
101	taehyeon1**	0.005079	142	Arken_K**	0.002711
102	sonju**	0.004715	143	buncoch**	0.002711
103	yonhaptv_e**	0.004715	144	skcc_t**	0.002697
104	would_you_**	0.004592	145	bstaeb**	0.002671
105	bhASAnti**	0.004592	146	Hansarangn**	0.002585
106	bluenote02**	0.004592	147	ooobbb**	0.002572
107	djuna**	0.004592	148	cns53**	0.002572
108	giantro**	0.004592	149	tnwkor**	0.002466
109	14O**	0.004589	150	maru51**	0.002445
110	casex90**	0.00437	151	booknp**	0.00242
111	Hongik**	0.004288	152	cosse01**	0.002411
112	kscmylife**	0.004288	153	sunkitty**	0.002337
113	BaobabDunc**	0.004208	154	blueballoon0**	0.002317
114	yonhaptwe**	0.004062	155	ksy445**	0.002286
115	CrowKing6**	0.00399	156	dambae_j**	0.002264
116	romnb**	0.003917	157	ellie**	0.002264
117	aromayoungk**	0.003917	158	Kyeongh**	0.002251
118	HuffPostKor**	0.003876	159	alleci**	0.002243
119	lovefu**	0.003751	160	iamkep**	0.002215
120	ktnbo**	0.003579	161	rlarudf**	0.00216
121	sadarl**	0.003466	162	minsu222**	0.002131

Rank	user_screen_name	KpRank 값
163	Seopomaanpi**	0.002131
164	poca_preghie**	0.001939
165	Arsenic**	0.001885
166	dhzon**	0.001848
167	TX_INSPIRATI**	0.001827
168	CodeGenerat**	0.00164
169	estim**	0.001581
170	laina9610**	0.001576
171	JSY_worldmus**	0.001537
172	hellosamy**	0.001519

A Study on Detection of Subject-based Key Player on Social Media Using KeyplayerRank

Kim, Minseon

Department of Civil and Environmental Engineering

The Graduate School

Seoul National University

Abstract

As the number of social media users steadily increases each year, more and more people are using social media to acquire data and information and communicate with other users on social media. When it comes to acquiring data or information, it often happens that people trust and follow the words of the most influential users in social media. Therefore, in this study, we want to find key players in social media by using KeyplayerRank applying Topical Similarity, which weights the relationship between people who are interested in a specific topic, and Influence Index considering user's influence and

power. In this study, the key player of the subject was detected on Twitter, LDA was used as a topic modeling method to extract the subject, and SVM was used to classify only text related to the topic. In this study, we proposed a new method to select users according to the purpose of finding key player in social media considering both indicators at the same time. The proposed methodology can be applied not only to Twitter but also to other social media, and it can also help policy decision through the marketing of the enterprise and public opinion formation.

In the future, this study can be developed as a research to find a real time key player when each topic becomes an issue in real time including spatio-temporal elements.

keywords : KeyplayerRank, PageRank, LDA, SVM, key player, social media

Student Number **: 2016-21243**