Journal of The Korea Big Data Service Society
Vol. 2. No. 1, June 2015, Pages 11-23

한국빅데이터서비스학회 논문지
ISSN 2383-9287

SNS 사용자의 감정 분석에 의한 영향력 측정

Influence Measurement based on Sentiment Analysis of SNS Users

정회윤(Hoe-Yun Jeong)[1], 지상훈(Sang-Hun Ji)[2], 양형정(Hyung-Jeong Yang)[3],
김경윤(Kyoung-Yun Kim)[4], 김경백(Kyung-Baek Kim)[5]

요 약

소셜 미디어의 등장으로 온라인상에서 정보 교류가 활발하게 이루어지고 있으며 소셜 미디어를 통한 여론형성, 의제설정 등과 같이 사회에서 일어나는 다양한 사건들에 큰 영향력을 발휘하고 있다. 본 논문에서는 소셜 미디어 중 하나인 트위터 상에서 큰 영향력을 발휘하는 영향력자(Influential) 또는 오피니언 리더(Opinion Leader)에 대한 영향력 측정을 제안한다. 기존의 영향력 측정 연구들은 팔로워(Follower), 리트윗(Retweet), 멘션(Mention)을 이용한 사용자 네트워크에서의 구조적인 요소를 통해 영향력을 측정 하였지만, 본 논문에서는 구조적인 요소뿐만 아니라 사용자들 간의 감정(Sentiment) 유사도 분석을 통해 영향력을 측정한다. 본 논문에서 제안하는 방법을 통해 선정된 영향력이 높은 사용자로부터 시작된 정보에 대해 네트워크상의 정보 확산 모델을 이용하여 영향력 최대화 문제에 적용함으로써, 기존의 영향력 측정 방법과 정보 확산 결과에 비교하였다. 이를 통해 본 논문에서 제안한 방법이 다른 영향력 측정 방법에 비해 높은 성능을 나타낸다는 것을 확인할 수 있었다. 또한, 이러한 결과를 통해서 감정적인 요소가 영향력 및 정보 확산에 많은 영향을 미친다는 것을 확인 할 수 있었다.

주제어: 빅데이터, 감정분석, 영향력, SNS, 영향력 최대화

1 전남대학교 전자컴퓨터공학과 석사 졸업
2 전남대학교 전자컴퓨터공학과 석사 과정
3 전남대학교 전자컴퓨터공학과 교수, 교신저자
4 웨인 주립대학교 산업공학과 교수
5 전남대학교 전자컴퓨터공학과 교수

Journal of The Korea Big Data Service Society
Vol. 2. No. 1, June 2015, Pages 11-23

한국빅데이터서비스학회 논문지
ISSN 2383-9287

Abstract

Measuring influence on social networks has attracted tremendous interest from both academia and industry. Social Network Services are known as an effective marketing platform where customers trust the advertisements which are provided by their friends and neighbors. Therefore, selecting seed user is the primary concern in viral marketing. In addition, most of the developed algorithms and tools mainly depend on the static network structure. In this paper, we propose influence measurement based on sentiment analysis in the social network. This model considers the most influential user in the community as the candidate for the top-k seeds. We employ influence maximization problem for evaluating proposed method. Experiments show that the proposed method performs consistently well in influence maximization.

Keywords: Big Data, Sentiment Analysis, Influence, Social Network Service, Influence Maximization

Journal of The Korea Big Data Service Society
Vol. 2. No. 1, June 2015, Pages 11-23
한국빅데이터서비스학회 논문지
ISSN 2383-9287

1. 서 론

현대 사회는 인터넷과 스마트폰의 보급으로 인해 사회 구성원들 간의 정보 교류가 끊임없이 이뤄지는 소셜 네트워크 사회(Social Network Community)이다. 이러한 사회로의 발전에는 스마트폰과 같은 인터넷 장비들뿐만 아니라 트위터(Twitter), 페이스북(Facebook), 인스타그램(Instagram) 등 다양한 소셜 미디어(Social Media)의 사용이 큰 역할을 차지하고 있다[14, 15, 18, 19]. 소셜 미디어의 사용이 증가함에 따라, 쉴 새 없이 공유되는 정보들이 끊임없이 축적되고 있다. 이러한 정보들이 여론형성, 의제설정 등에 사용되면서, 소셜 미디어의 정보 전파에 대한 사용자의 영향력 측정 연구가 증대되고 있다.

영향력 측정 및 분석은 마케팅, 정치, 광고 등의 다양한 분야에서 중요한 역할을 차지한다. 마케팅 분야에서, 영향력은 제품을 홍보하고 평판을 유도하는 역할을 한다. 정치가들에게는 홍보 및 선거의 승패를 예측하는 중요한 요소로 작용한다. 또한, 광고 분야에서 영향력 측정은 빠른 정보전달과 저비용 고효율의 정보 전달 측면에서 중요하다.

기존의 많은 영향력 측정 방법들은 SNS상의 구조적인 관점에서 바라본다. 트위터를 이용한 영향력 측정의 경우, 사용자 간의 팔로워(Follower), 리트윗(Retweet), 답글(Reply), 멘션(Mention) 등과 같은 트위터의 구조상 정보만을 이용하여 사용자의 영향력을 예측했다[1, 2, 3]. 그러나 감정적인 단어에 대해 확산 속도가 다르다는 연구 결과를 통해 감정적인 요소들이 정보 전달, 홍보, 여론형성, 의제설정과 같이 의사결정에 영향을 준다는 것을 알 수 있다. 이를 통해 감정적인 요소와 정보 확산 영향력이 밀접한 연관이 있다고 할 수 있다[4, 5].

본 논문에서는 영향력 측정을 위해 팔로워, 리트윗, 답글, 멘션 등의 구조적인 요소들을 통해 트윗의 전달 확률을 계산하고 사용자들의 감정적인 유사도가 높을수록 정보 전달에 대해 높은 가중치를 적용한다. 이렇게 측정된 영향력에 대해 감정적 요소들을 고려함으로써 정보들이 어떻게 전달되는지를 분석한다. 또한, 트위터 사용자들의 감정을 고려한 요소들이 정보전달에 얼마나 영향을 주는지에 대해 분석한다. 사용자간의 감정 요소가 정보 전달에 영향을 미치는지 검증하기 위해 정보 전달 확산 모델인 영향력 최대화(Influence Maximization)를 통해 기존의 영향력 측정 방법들과의 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 영향력 측정과 관련된 연구들을 살펴보고, 3장에서는 트위터에서 구조적인 요소와 감정적인 요소가 병합되어 정보 전달에 대한 영향력을 측정하는 방법을 제시한다. 4장에서는 실험을 통해 감정요소가 정보전달에 미치는 영향력에 대한 실험 결과를 살펴보고 5장에서 결론을 제시한다.

2. 관련 연구

영향력을 측정하는 대표적인 연구에는 중앙성(Centrality)을 이용한 영향력 측정 방법이 있다. 중앙성을 측정하는 방법에는 이접 중앙성, 매개 중앙성, 연결 중앙성 등의 다양한 방법들이 존재한다. 그 중 이접 중앙성(Closeness Centrality)[8, 9, 10]은 서로 다른 두 사용자 간의 최단 경로를 측정하여 최단경로들의 합이 가장 작은 사용자를 전체 네트워크

에서 가장 영향력이 높은 사용자로 분류한다. [11]는 매개 중앙성(Betweenness Centrality)을 적용하여 네트워크상에서의 한 사용자가 다른 사용자들 사이에 위치하는 정도를 정의하는 것으로, 한 점이 담당하는 중재자 역할의 정도로써 중앙성을 측정한다. 즉 최단 경로 위에 위치하면 할수록 그 사용자의 영향력은 높아진다. 연결 중앙성(Degree Centrality)은 다른 점과 연결된 정도를 중시하며, 연결망 내에서 한 점에 연결되어 있는 점들의 합을 말한다. 영향력은 전체 연결 수에서 각 행위자의 내향 연결 정도와 외향 연결 정도의 비율로 측정된다. 이러한 구조적인 중앙성을 이용한 영향력 측정은 구조만을 고려하는 한계를 벗어나지 못한다는 문제점이 있다. 즉, 사용자들의 상관관계를 단순히 구조적인 특징만을 통해서 영향력을 측정한다는 점에서 다양한 사회적 관계성을 표현해야 하는 현실 세계의 문제를 제대로 표현하기 힘들다는 단점이 있다.

또 다른 대표적인 영향력 측정 알고리즘으로는 Google의 검색에 적용된 페이지랭크(PageRank) 알고리즘이 있다[12]. 페이지랭크는 상대적 중요도에 따라 가중치를 부여하여 영향력을 측정한다. 페이지랭크는 서로의 인용과 참조로 연결된 임의 그룹에 적용된다. 하지만 현실세계에 존재하는 소셜 네트워크 특성상, 다수의 의견보다 소수의 의견이 더 높은 영향력을 제공할 수 있으므로 상대적인 중요성이 높다고 해서 영향력이 높다고 할 수는 없다.

[13]은 'SNS 상에서 팔로어가 많은 사람이 영향력이 있다'라는 가설에 반문하기 위해 시작된 연구이다. 이 연구에서는 과학적인 데이터 기반의 검증을 위해 2009년 트위터의 전체 데이터를 수집했다. 사용자의 영향력 척도로서 팔로어의 수, 트윗에 대한 대답 횟수, 리트윗의 수를 측정하였다. 세 가지의 척도를 이용하여 영향력을 측정한 결과, 영향력의 순위는 일정하지 않고 척도에 따라 다르게 나타나는 것을 확인 할 수 있었다. 데이터 분석 결과, 팔로어와 언급 혹은 리트윗과의 관계는 낮은 상관관계를 보였다. 즉, 팔로어가 많은 인기 있는 트위터 사용자라고 반드시 언급이나 재전송이 많이 되지는 않는다는 것이다. 즉, 인기와 영향력은 다르다는 것이다. 따라서 이러한 문제를 해결하기 위해 사용자간의 감정적 유사성을 고려한 영향력 측정 연구가 필요하다.

3. 제안방법

본 장에서는 트위터 상에서의 정보 전달의 영향력을 측정하기 위한 방법을 제안한다. 영향력 측정을 위해, 본 논문에서는 트위터 데이터를 수집하고, 자연어 처리 및 감정분석을 수행한다. 분석된 정보들은 사용자의 트윗에서 다른 트윗으로 전달 될 확률 계산에 이용되며, 이러한 트윗의 전달될 확률 정보를 바탕으로 사용자간의 리트윗, 팔로워, 답글, 멘션 등을 고려하여 상대적 중요도를 분석하고, 감정적 유사도 가중치를 적용하여 영향력을 파악한다.

3.1 전처리 단계

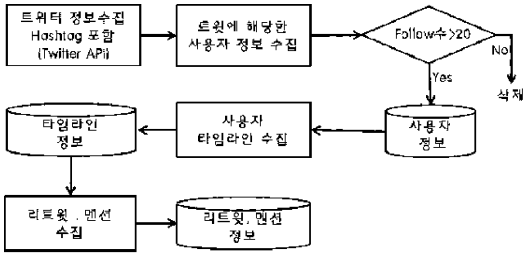

그림 3.1 트위터 데이터 수집 구성도

데이터 수집은 [그림 3.1]에서와 같이 트위터 데

이터를 수집 및 수집된 데이터를 타임라인과 리트윗, 멘션으로 분리한다. 트위터 상의 데이터는 REST API를 이용하여 수집한다[6]. REST API는 과거에 발생한 트윗에 접근할 수 있기 때문에 정의할 수 있는 검색 조건의 종류가 다양하다. REST API를 이용하여 수집된 정보들은 XML(확장성 생성언어) 파일로 된 웹페이지를 읽어서 원하는 정보를 수집하는 구조로 이루어졌으며, 사용자의 타임라인, 팔로어, 리스트, 유저 정보, Account, Location, Geo 등의 정보를 검색한다. 본 논문에서는 트위터상의 사용자의 영향력을 측정하기 위해서 특정 기간 동안 인기를 끌었던 토픽에 대한 트위터 데이터를 수집한다. 수집된 정보를 바탕으로 각 항목에 대한 리트윗, 답글, 팔로워와 멘션 등에 대해 네트워크를 구축한다.

3.2 자연어 처리

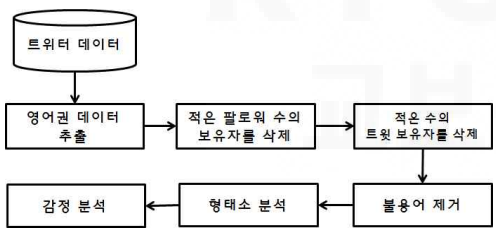

그림 3.2 데이터 자연어 처리

그림 3.2는 수집된 트위터 데이터를 이용하여 자연어 처리를 하는 과정을 나타낸다. 영어권 데이터를 사용하기 위해 미국 지역 내에서 발생하고, 영어로 작성된 트윗만을 필터링한다. 이는 영어권에서 사용하는 데이터의 수와 사용자간의 관계가 다양하기 때문이다. 또한, 영향력 측정에 비중이 적은 팔로워 수와 트윗 작성 수가 적은 사용자에 대한 데이터는 제외한다. 트위터 데이터의 자연어 전처리를 위하여, 각 사용자가 보유한 트윗 정보에 대해서는 어근 분석(Stemming), 어근 추출, 단어 분리 과정 등의 일반적인 용어 추출에 대한 전처리 과정을 수행한다. 이 과정에서 사용자별 용어 출현 빈도수는 해쉬 태그를 기준으로 감정 분석을 위하여 추가적으로 저장한다.

불용어는 관사나 전치사, 조사, 접속사와 같이 검색을 할 때 의미가 없는 단어들을 뜻한다. 본 논문에서는 금칙어를 제외한 불용어 테이블을 이용하여 불용어를 삭제하였다. 어근 분석 작업은 예를 들어 명사의 경우 단수 형태 ‘apple’, 복수 형태 ‘apples’ 같이 동일한 단어임에도 수에 따라서 형태가 다른데 이를 원형으로 변환하는 과정을 의미한다. 위의 ‘apple’과 ‘apples’의 경우, 변환 과정을 거쳐 기본형 ‘apple’로 변환 될 것이다. 해당 과정을 통해 트윗 집합으로부터 용어 및 출현 빈도로 구성된 테이블을 만든다.

3.3 감정분석

트윗(Tweet)에 포함된 감정 분석을 위해, 트위터 사용자가 트윗 작성 시 어떤 감정을 트윗에 담고 있는지를 파악해야 한다. 사용자가 어떤 감정을 트윗에 담았느냐에 따라서 트윗의 특성이 달라진다. 트윗을 작성하거나 리트윗에 의견을 덧붙인 사용자가 부정적인 평가를 했다면 이 트윗이 전파되는 과정에서 부정적인 분위기로 인한 선입견이 만들어져 트윗을 보는 사람들에게 부정적인 영향을 미칠 수 있다. 본 논문에서는 실질적인 감정의 전파 과정을 분석하기 위해 트윗에 내재된 사용자의 감정을 추출하고 트윗에 드러난 감정적 값을 추출한다.

본 절에서는 전처리 단계에서 트윗을 형태소 단위로 분리하고, 분리된 정보를 기반으로 미리 정의 되

어 있는 감정 리스트에 포함된 트윗만을 추출하여 레이블링(Labeling)을 한다. 수집된 트윗들은 일상생활에서 사용하는 자연어 형태의 언어들이기 때문에 컴퓨터가 처리할 수 있도록 자연어 처리(NLP, Natural Language Processing) 과정이 요구된다. 감정 분석을 위해 극성 레이블을 구성하여 트윗 데이터에서 추출된 키워드에 적용한다. 본 논문에서는 자연어 처리를 통해 수집한 트윗에서 극성을 갖는 모든 성상형용사, 상태성명사를 추출하여 극성 레이블을 부여한다.

본 논문에서는 Stanford Sentiment Treebank를 이용하여 감정을 분석한다[7]. 이 방법은 Recursive Neural Tensor Network을 이용하여 각 단어들을 Tree 구조로 표현하고, 미리 정의된 감정 사전을 이용하여 각각의 감정값들을 매핑 시키는 방법이다. 본 논문에서는 Stanford에서 제공하는 감정 카테고리를 가진 API를 이용하여 감정 정보를 구분하고 각 카테고리 마다 나타나는 감정의 정도를 분석한다.

3.4 트위터 모델링

트위터에서는 다양한 방법을 이용하여 트위터 사용자들끼리 정보를 교환하는 행위들이 존재한다. 사용자들은 다른 사용자와의 팔로워(Follower) 관계를 통해서 팔로워한 사용자의 트윗을 확인하거나, 팔로워한 사용자가 다른 사용자의 트윗을 리트윗(Retweet)하거나, 혹은 답글을 작성함으로써 팔로워가 아닌 제 3의 사용자가 그 트윗을 전달 받을 수 있다.

본 절에서는 트위터 상 사용자의 영향력을 나타내기 위해 트위터에 대해서 다음과 같이 모델링하였다. 수집 된 데이터에서, 트윗들의 집합은 T, 사용자 집합을 U, 해쉬태그들의 집합은 HS 라고 정의한다.

리트윗(Retweet, RT)은 트윗 i 를 리트윗하여 다른 사용자의 타임라인에 트윗 j 가 작성된 경우로 식 (1)과 같이 나타낸다.

$$T_{i,j} \begin{cases} 1 & tweet\ i\ reteets\ tweet\ j \\ 0 & otherwise \end{cases}, \quad RT = |T| \times |T| \tag{1}$$

답글(Reply, RP)은 트윗 i 에서 어떤 사용자가 답글을 작성하여 트윗 j 가 생성된 경우로서 식 (2)와 같다.

$$RP_{i,j} = \begin{cases} 1 & tweet\ i\ replies\ tweet\ j \\ 0 & otherwise \end{cases}, \quad RP = |T| \times |T| \tag{2}$$

멘션(Mention, MN)은 사용자 j 에게 멘션으로 트윗 i 를 작성한 경우로 식 (3)과 같다.

$$MN_{i,j} = \begin{cases} 1 & tweet\ i\ mentions\ user\ j \\ 0 & otherwise \end{cases}, \quad MN = |T| \times |U| \tag{3}$$

팔로잉(Following, FW)은 사용자 i 가 사용자 j 를 팔로우한 경우이며, 이에 대해 식 (4)와 같이 나타낸다.

$$FW_{i,j} = \begin{cases} 1 & user\ i\ follows\ user\ j \\ 0 & otherwise \end{cases}, \quad FW = |U| \times |U| \tag{4}$$

해쉬태그(HashTag, HT)는 트윗 i 에서 해쉬태그(HS)를 이용하여 작성한 경우이고 식 (5)와 같다.

$$T_{i,j} \begin{cases} 1 & tweet\ i\ includes\ hashtag\ j \\ 0 & otherwise \end{cases}, \quad HT = |T| \times |HS| \tag{5}$$

3.5 트위터 영향력 측정

트위터에서는 사용자가 작성한 트윗이 리트윗, 멘션, 답글을 통해서 주변의 사용자에게 전달된다. 하지만, 트위터의 각 트윗간의 희박한 연결성을 가진 구조로 인해 직접적으로 링크를 통해 전달되는 행동인 리트윗, 멘션, 답글과 같은 행동만을 고려한다면 영향력을 정확하게 계산하기 어렵다. 본 논문에서는 트윗들의 관계를 고려할 뿐만 아니라 트윗들의 전달되는 방법에 대해서도 고려하여 영향력을 측정한다.

$$,j \quad \alpha R_{i,j}+\beta L_{i,j}+\gamma M_{i,j}+\delta F_{i,j}+\epsilon H_{i,j}, \quad (\alpha+\beta+\gamma+\delta+\epsilon=1) \tag{6}$$

식 (6)은 감정적인 요소를 고려하지 않고 구조적인 요소만을 고려한 영향력 측정 계산 방법이다. 식 (6)을 통해서 사용자의 행동을 통해 정보가 전달될 확률을 정의한다.

수식 (6)에서 G는 랜덤하게 트윗에 접근할 확률과 트위터 상에서의 활동을 통해 접근 할 수 있는 경우로 나눌 수 있다. 또한, , $\beta, \gamma, \delta, \epsilon$의 값은 영향력 점수에서 $\alpha, \beta, \gamma, \delta, \epsilon$ 각각에 가중치를 부여해 준다. 이들의 합은 1로 한정 지으며, 이는 각 요소의 가중치 값이 너무 커지거나 작아지는 것을 방지하기 위함이다. 각 트위터 상의 활동을 통해 얻어지는 확률 계산은 다음과 같다.

사용자가 작성한 트윗은 리트윗이나 답변을 통해 다른 트윗으로 전달된다. 식 (1)과 (2)를 이용하여, 트윗 i에서 리트윗이나 답변을 통해 트윗 j로 전달될 확률($_{,j}$)을 식 (7)과 같이 정의한다.

$$L_{i,j}=RT_{i,j}+RP_{i,j} \tag{7}$$

사용자 i가 작성한 전체 사용자에 대한 멘션을 FM이라고 할 때, 이 중에 사용자 j에게 작성한 확률에서 선택한 사용자 j의 사용자에게 선택될 확률($M_{i,j}$)과 같다.

$$FM_{i,j}=\begin{cases} MN_{i,j} \\ m_i \\ 0 \end{cases}, \begin{matrix} m_i>0 \\ otherwise \end{matrix} \tag{8}$$

m_i 는 트윗에서 멘션을 가지고 있는 트윗의 수이다.

$$M_{i,j}=\frac{FM_{i,u}}{|t \quad T:u_t=u_j|} \tag{9}$$

$|t \in T:u_t=u_j|$는 u_j의 전체 트윗 수이다.

사용자가 i의 팔로워들을 통해서 사용자 j로 접근할 확률이 $FW_{i,j}$ 라고 하면 이러한 접근을 통해서 사용자 j의 트윗에 접근 할 수 있다. 트윗 i가 사용자 i의 팔로워 정보를 통해서 사용자 j로 접근해서 트윗 j로 접근할 확률($F_{i,j}$)은 다음과 같이 나타낼 수 있다.

$$FF_{i,j}=\begin{cases} \frac{FW_{i,j}}{f_i} \\ 0 \end{cases}, \begin{matrix} f_{i>0} \\ otherwise \end{matrix} \tag{10}$$

는 사용자 i의 팔로워 수이다.

$$F_{i,j}=\frac{FF_{u_i,u_j}}{|t \in T:u_{t=}u_j|} \tag{11}$$

$|t \in T:u_t=u_j|$는 u_j의 전체 트윗 수이다.

트윗간의 공통 주제를 가진 사용자일수록 서로 다른 주제를 가진 사용자보다 접근을 자주한다. 사용

자 i, j가 각각 유사한 주제를 가지고 있는지를 정의하기 위한 식($_{,j}$)은 다음과 같다.

$$P_{,j,k} \begin{cases} HT_{j,k} \\ ht_k \\ 0 \end{cases} \bullet \left(\frac{HT_{i,k}}{th_i} \right), ht_k > 0 \text{ and } th_i > 0, otherwise \quad (12)$$

t 는 해쉬태그 k를 포함한 트윗의 수이고, th_i는 트윗 i 에 포함된 전체 해쉬태그의 수이다.

$$_{i,j} = \underset{k}{\underset{S}{HP_{i,j,k}}} \quad (13)$$

사용자 i 와 사용자 j 의 감정적 유사성이 비슷할 때 전달될 확률($S_{i,j}$)은 사용자들의 공통 해쉬태그를 이용하여 사용자들이 한 주제에 대해서 어떠한 감정을 가지는 가를 정의한다. 사용자 i 와 j 가 공통적으로 가지고 있는 해쉬태그에 대한 감정 분석 결과 집합을 hs라고 할 때, 사용자간의 감정의 유사도는 코사인 유사도를 통해 식 (14)와 같이 정의한다.

$$S_{i,j} = \frac{hs_i \bullet hs_j}{|hs_i| \bullet |hs_j|} \quad (14)$$

식 (14)를 이용하여 식 (6)에서 정의한 영향력 측정 방법에 감정에 대한 가중치를 식 (15)와 같이 적용한다.

$$Z = \frac{\beta}{1-\alpha} L_{i,j} + \frac{\gamma}{1-\alpha} M_{i,j} + \frac{\delta}{1-\alpha} F_{i,j} + \frac{\epsilon}{1-\alpha} H_{i,j} \quad (15)$$

영향력 G는 다음과 같다.

$$G^{\cdot j} = \alpha R_{i,j} + (1-\alpha) Z \quad (16)$$

식 (16)에서 사용자 i 와 j 의 감정적 유사함에 영향을 미치는 리트윗, 멘션, 답글은 Z 에 감정적 유사함에 대한 가중치를 적용한다. 감정적 유사도가 사용자의 행동에 대한 전달 확률에 영향을 미치기 때문에 Z 에 한해서만 가중치를 적용한다. 감정적 가중치를 적용한 방법에 대해서는 식 (17)과 같이 정의한다. 사용자 i, j 에 대해서 각각의 주제에 대한 서로의 가중치 정보를 분석한다.

$$Z'^{i,j} = (1-\alpha) \frac{(S_{i,j}+k)}{\sum_{u \in U}(S_{i,u}+k)} Z_{i,j} \quad (17)$$

상수 k는 유사도 결과가 0일 경우 발생하는 결과에 대해서 잘못된 결과를 나타내는 것을 방지하기 위해서 0보다 큰 임의 값으로 한다. 최종적으로 감정을 고려한 영향력G'는 식(18)과 같다.

$$G' = \alpha R + (1-\alpha) Z' \quad (18)$$

4. 실험

4.1 데이터 수집

본 논문에서는 2014년 4월부터 트위터 REST API를 이용하여 트윗 데이터를 수집하였으며, 수집된 데이터에는 특정 토픽을 포함한 데이터와 그 데이터를 게시한 사용자들의 정보를 포함하였다. 수집된 정보에서는 영어권 사용자들만을 대상으로 하였고, 비교적 영향력이 적다고 판단되는 팔로워의 수가 20 이하인 사용자들을 제외하였다. 본 논문에서는 수집된 데이터를 이용하여 트윗들에 대한 감정 분석을 실시하였고, 사용자들의 감정 유사도를 계산하였다. 본 논문에서는 특정 토픽 4개(Galaxy, iPhone, Android, iOS)를 이용하여 데이터를 수집하

였다. 이렇게 수집된 데이터를 각각 T1, T2, T3, T4로 명명하여 실험을 진행하였다.

	T1	T2	T3	T4
Number of users	11K	10K	9K	11K
Number of tweets	77K	12K	25K	19K
Number of follower links	47K	45K	32K	37K
Average Degree	14.2	15.2	11.9	20.5

표 4.1 수집된 데이터 집합

4.2 영향력 최대화

영향력 최대화 문제의 정의는 다음과 같이 주어진 그래프 (V,E)를 사용자 V와 그들의 각 관계 E로 표현한다. 이에 대해 각 링크 (u,v)에서 전파 확률 $p_{,v}$ $[0,1]$이 할당된다고 할 때, 영향력 최대화 문제는 영향력 함수 $f(S)$를 최대화하는 k개의 노드들로 이루어진 부분집합 $S \subseteq V$를 선택하는 것이다. 이때, 영향력 함수 $f(S)$를 선형 임계값 모델(Linear Threshold Model, LT-Model)과 독립 캐스케이드 모델(Independent Cascade Model, IC-Model)을 통해 정보 전파될 때 정보가 전달되는 노드 개수에 대한 예측값으로 정의한다[16, 17, 20, 21].

4.3 비교 알고리즘

본 논문에서는 트윗 속에 존재하는 감정적인 요소가 최초 정보 전파시 얼마만큼의 전파력을 가지고 있는지를 통해 사용자들의 영향력 측정을 평가한다. 이를 위해 영향력이 높다고 측정된 k명의 사용자를 통해 메시지가 얼마나 많이 전달되는지 확인 할 수 있는 영향력 최대화 결과를 통해 본 논문에서 제안한 영향력 측정 방법과 비교한다. 1~50개의 k의 값을 변화시켜 실험을 진행하였다.

본 논문에서는 영향력 평가를 위해 제안한 방법과 기존의 영향력 측정 방법(중앙성, 페이지랭크)들을 비교하였다. 또한, 제안한 방법과 제안한 방법에서 감정 가중치를 고려하지 않은 상태에서 영향력 최대화 방법을 비교하였다. 영향력 최대화에 대해 정확한 실험을 위해, 각각의 알고리즘들을 여러 번 반복하여 평균결과를 계산한다.

4.4. 실험 결과

다음 그림 4.1, 4.2는 기존 알고리즘과 본 논문에서 제안한 방법을 이용해 k 값(사용자의 수)의 증가에 따라 변화하는 최대 영향력을 측정한 결과이다. 이를 통해 시드 k의 값이 높을수록 정보 확산이 증가함을 확인 할 수 있으며, 그 중 본 논문에서 제안된 방법은 같은 k 값에서 기존 방법들 보다 더 높은 정보 확산 능력을 가짐을 확인 할 수 있었다. 이를 통해 정보 확산 능력이 높을수록 영향력이 높음을 확인할 수 있다.

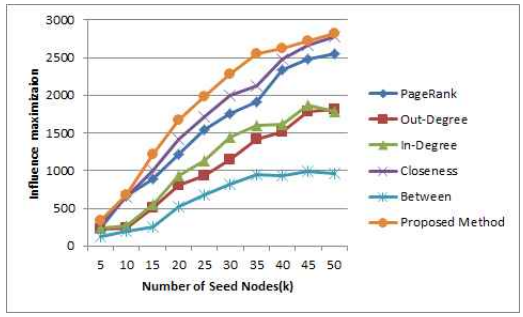

그림 4.1 T3에서의 영향력 최대화에 대한 성능

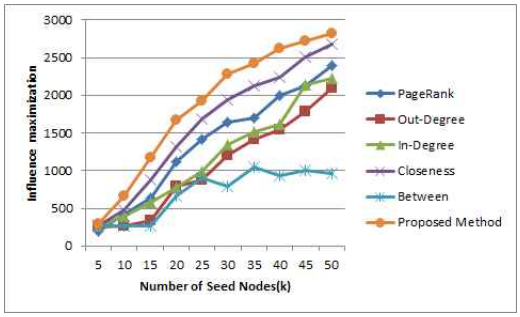

그림 4.2 T4에서의 영향력 최대화에 대한 성능

또한, 표 4.2는 시드 k가 50일 때, 각 메시지의 확산 정도를 나타난 결과이다. 다른 비교 모델을 통해서 본 논문에서 제안하는 방법이 높은 확산을 보이는 것을 확인할 수 있다. 이를 통해 영향력 측정에서 구조적인 측면만을 고려한 방법보다는 감정적인 요소가 고려된 정보를 통해 영향력이 높은 사용자를 측정함으로써 그 사용자가 높은 정보 확산 능력을 가지고 있음을 확인 할 수 있다.

	T1	T2	T3	T4
PageRank	1522	2417.4	2555	2396
Out-Degree	1255	1987	1811	2100
In-Degree	1333	2096.1	1788	2230
Closeness	1634	2701.3	2777	2671
Between	889	1132	899	965
Proposed method	1689	2798	2800	2944

표 4.2 k가 50일 때, 영향력 최대화를 통한 결과 비교

또한, 본 논문에서 감정적인 요소를 고려하지 않은 실험 결과와 비교를 통해 감정적인 요소가 영향력에 미치는 범위를 명확히 확인 할 수 있다. 감정적인 요소를 고려한 그림 4.3과 4.4의 그래프(Sentiment)의 값이 감정적인 요소를 고려하지 않은 그래프(Non-Sentiment)의 값보다 더욱 높은 영향력을 보임을 확인 할 수 있다.

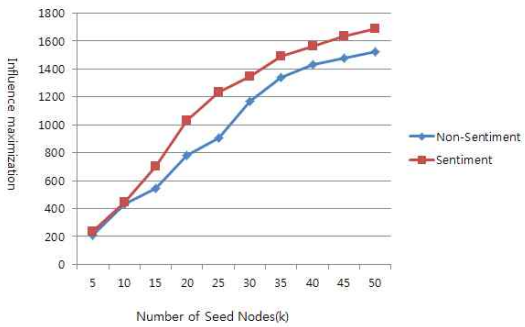

그림 4.3 T1에서의 영향력 최대화에 대한 성능

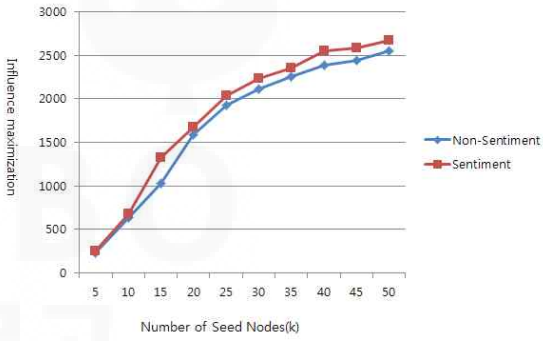

그림 4.4 T2에서의 영향력 최대화에 대한 성능

표 4.3은 k = 50에 대해서 영향력 최대화 알고리즘을 적용했을 때, 영향력 최대값에 대한 결과를 표로 나타낸 것이다. 이를 통해서 감정적인 요소를 고려한 방법에 대해서 평균적으로 114.8% 상승함을 확인 할 수 있다.

	T1	T2	T3	T4	T5
Sentiment	1758	2818	2671	1689	2554
Non-Sentiment	1522	2300	2396	1522	2244
증가율	115.5 %	122.5 %	111.4 %	110.9 %	113.8 %

표 4.3 영향력 최대화를 통한 결과 비교

5. 결론 및 향후 연구

본 논문에서는 트위터 사용자가 트윗을 작성했을 때, 그 트윗이 사용자들의 감정적 유사도를 바탕으로 특정 행동(retweet, mention, reply)을 통해서 얼마나 많은 인원들에게 전달되는지의 정도를 영향력으로 나타내었다. 이러한 영향력에서 단순히 구조적인 요소를 고려하는 것 외에도 사용자들이 작성한 트윗들의 감정적인 요소들을 통해 영향력의 전파에 영향을 미치는지를 확인하였다. 이를 위해 트위터 정보를 이용하여 트위터 상의 사용자 관계 및 사용자의 행동들의 관계를 분석하고 사용자와의 감정적 유사 정도에 따라 영향력에 대한 가중치를 적용하여 영향력 측정하였다.

측정된 사용자들의 영향력을 검증하기 위해 네트워크상의 정보 확산 모델을 이용하여 영향력 최대화 문제에 적용함으로써, 기존의 영향력 측정 방법과의 정보 확산 결과와 비교하였다. 이를 통해 본 논문에서 제안한 방법에 대한 성능이 우수함을 확인하였다. 또한, 감정을 고려하지 않은 페이지 랭크 알고리즘과 비교하였을 때, 더 높은 정보 전파력을 나타냈으며, 이러한 결과를 통해서 감정적인 요소가 영향력 및 정보 확산에 많은 영향을 미친다는 것을 확인할 수 있었다.

향후 연구로는 감정적인 요소를 고려한 영향력 측정 외에도 영향력 측정과 관련된 다른 요소들을 영향력 측정에 적용하고 이를 기존의 연구 방법과 비교하는 것을 제안한다.

6. 참고 문헌

[1] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. "Influence and passivity in social media". In 20th International World Wide Web Conference (WWW'11), 2011.

[2] Auvinen, Ari-Matti. Social media - the new power of political influence. Centre for European Studies. (2011)

[3] Wei Chen , Yajun Wang , Siyu Yang, Efficient influence maximization in social networks, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 28-July 01, 2009, Paris, France

[4] 이승희, 박영호, "소셜 네트웍 영향력 측정 모델 제안", 한국정보처리학회 2011년도 제35회 춘계학술발표대회

[5] 최준일 "클러스터링 및 랭킹 기법을 적용한 트위터 사용자의 영향력 측정에 관한 연구", 대구대학교 학위논문 2013

[6] https://dev.twitter.com/rest/public

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP 2013, pages 1631-1642.

[8] S. Hakimi. Optimum locations of switching

centers and the absolute centers and medians of a graph. Operations Research, 12:450{459, 1965.

[9] G. Sabidussi. The centrality index of a graph.Psychmetrika, 31:581−603, 1966.

[10] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge Univ.Press, 1994.

[11] Shamanth Kumar, Fred Morstattter, Huan Liu, "Twitter Data Analytics", Springer, August 19, 2013

[12] 박지혜, 서보밀, “온라인 소셜 네트워크 서비스 환경에서 유력자의 매개 중심성이 구전 효과에 미치는 영향”, Journal of information technology applications & management

[13] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, Ed H. Chi, and Rowan Nairn, "Short and Tweet: Experiments on Recommending Content from Information Streams," in Proc. of the 28th international conference on Human factors in computing systems (CHI '10), pp.1185−1194, 2010.

[14] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake Shakes Twitter Users: Realtime Event Detection by Social Sensors," in Proc. of the 19th international conference on World wide web (WWW '10), pp.851−860, 2010.

[15] 이미영, 최 완, "빅데이터 분석을 위한 빅데이터 처리 기술 동향", 정보처리학회지 제 19권 제 2호 p. 20−28, 2012. 3

[16] X. He, G. Song, W. Chen, and Q. Jiang. Influence blocking maximization in social networks under the competitive linear threshold model. In SDM, 2012.

[17] Amit Goyal , Wei Lu , Laks V. S. Lakshmanan, SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model, Proceedings of the 2011 IEEE 11th International Conference on Data Mining, p.211−220, December 11−14, 2011

[18] 정기주, 서효영, 조성도, “소셜 네트워킹 서비스(SNS) 관련 연구의 분류와 연구 동향”, 한국지식정보기술학회 논문지 제6권 제5호 2010년 10

[19] Nicola Barbieri , Francesco Bonchi , Giuseppe Manco, Topic−Aware Social Influence Propagation Models, Proceedings of the 2012 IEEE 12th International Conference on Data Mining, p.81−90, December 10−13, 2012

[20] Manuel Gomez−Rodriguez , Jure Leskovec , Andreas Krause, Inferring Networks of Diffusion and Influence, ACM Transactions on Knowledge Discovery from Data (TKDD), v.5 n.4, p.1−37, February 2012

[21] Seth A. Myers , Jure Leskovec, Clash of the Contagions: Cooperation and Competition in Information Diffusion, Proceedings of the 2012 IEEE 12th International Conference on Data Mining, p.539−548, December 10−13, 2012

정 회 윤

2013년 전남대학교 전자컴퓨터공학과 졸업(학사)
2015년 전남대학교 전자컴퓨터공학과 졸업(석사)
관심분야 : 데이터마이닝

지 상 훈

2011년 조선대학교 과학교육학부(물리교육) 졸업(학사)
2014년 - 현재 전남대학교 전자컴퓨터공학과 석사과정
관심분야 : 빅데이터

양 형 정

1991년 전북대학교 전산통계학과 졸업(학사)
1993년 전북대학교 전산통계학과 졸업(석사)
1998년 전북대학교 전산통계학과 졸업(박사)
2003년 - 2005년 카네기멜런 대학교 연구원
2005년 - 현재 전남대학교 전자컴퓨터공학부 조교수
2007년 - 현재 전남대학교 전자컴퓨터공학부 부교수
관심분야 : 데이터마이닝, 멀티미디어데이터분석, e-Design

김 경 윤

1996년 전북대학교 산업공학과 졸업(학사)
1998년 전북대학교 산업공학과 졸업(석사)
2003년 피츠버그대학교 산업공학 졸업(박사)
2003년 - 2005년 피츠버그대학교 연구교수
2005년 - 현재 Wayne State University 교수
관심분야 : 협업적설계 CAD/CAM, Telerehabilitation

김 경 백

1999년 한국과학기술원 전자전산(학사)
2001년 한국과학기술원 전자전산(석사)
2007년 한국과학기술원 전자전산(박사)
2007년 - 2011년 University of California, Irvine, 박사후연구원
2012년 - 현재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 분산시스템, 미들웨어, 피어투피어 네트워크, 오버레이