International Journal of Software Engineering
and Knowledge Engineering
Vol. 27, No. 5 (2017) 841–867

DOI: 10.1142/S0218194017500310

An Evaluation Method for Content Analysis Based on Twitter Content Influence

Euijong Lee[*,‡], Young-Gab Kim[†,§,‡‡], Young-Duk Seo[*,¶], Kwangsoo Seol[*,‖] and Doo-Kwon Baik[*,**,‡‡]

[*]*Department of Computer Science and Engineering*
Korea University, Seoul 136-713, Republic of Korea

[†]*Department of Computer and Information Security*
Sejong University, Seoul 143-747, Republic of Korea
[‡]*kongjjagae@korea.ac.kr*
[§]*alwaysgabi@sejong.ac.kr*
[¶]*seoyoungd@korea.ac.kr*
[‖]*seolks@korea.ac.kr*
[**]*baikdk@korea.ac.kr*

Received 3 April 2015
Revised 22 October 2015
Accepted 9 February 2017

Twitter is a microblogging website, which has different characteristics from any other social networking service (SNS) in that it has one-directional relationships between users with short posts of less than 140 characters. These characteristics make Twitter not only a social network but also a news media. In addition, Twitter posts have been used and analyzed in various fields such as marketing, prediction of presidential elections, and requirement analysis. With an increase in Twitter usage, we need a more effective method to analyze Twitter content. In this paper, we propose a method for content analysis based on the influence of Twitter content. For measuring Twitter influence, we use the number of followers of the content author, retweet count, and currency of time. We perform experiments to compare the proposed method, frequency, numerical statistics, user influence, and sentiment score. The results show that the proposed method is slightly better than the other methods. In addition, we discuss Twitter characteristics and a method for an effective analysis of Twitter content.

Keywords: Twitter; content influence; retweet; follower.

1. Introduction

Twitter is a microblogging website that allows users to create content that is less than 140 characters in length (called a tweet). This service also supports a one-directional relationship between users: a user can see another user's contents simply

‡‡Corresponding authors.

by being his/her follower. Because of these characteristics, sometimes, Twitter can spread news faster than news media. For example, the news about Hudson River crash landing and the death of Michael Jackson was first spread through Twitter. Currently, Twitter is not only a social networking service (SNS) but also a news media [9]. Further, Twitter content contains various types of information such as a user's thoughts, interests, and feelings. Moreover, tweets contain implicit information such as user-specific patterns, public concerns, and social trends. Therefore, Twitter information can be analyzed for the prediction of presidential elections, marketing information, and requirement analysis. With an increase in the number of Twitter application fields, the demand for effective analysis methods for Twitter content has increased.

The motivation for this research is that there are many brilliant methods to analyze Twitter content, but these methods require a considerably large amount of data for the analysis of Twitter content and the related preconditioning processes [1–3, 6, 14, 15, 17, 18, 20–22]. Therefore, for analyzing Twitter content, we propose a method using the characteristics of a single Twitter post. These characteristics include the number of retweets, the number of followers, and the posting time. In Twitter, such information can be extracted from a single tweet. Further, in our previous paper, we proposed a method for measuring content influence [10]. In this previous research, we defined the influence of Twitter content as the value of posts, and the method to compute the value of the content by using the number of retweets, the number of followers, and the posting time. In this paper, we propose a new approach based on our previous research [10] for analyzing trends. We performed experiments using real Twitter data and demonstrated that the method is useful for Twitter search. In the current study, we expand our previous method to analyze trends with a small amount of data. Further, we investigate characteristics that are useful for analyzing Twitter content. In this study, we perform an experiment to show the proposed method's excellence with real Twitter data and discuss the characteristics that are helpful for analyzing Twitter content.

The remainder of this paper is organized as follows: In Sec. 2, we present the related works. Section 3 introduces our previous research briefly and discusses the framework of the experimental environment and the proposed method. In Sec. 4, we describe our experiment and present the results. We discuss the features of Twitter content and a comparison with other methods in Sec. 5 and conclude the paper in Sec. 6.

2. Related Works

Twitter content research consists of an analysis of the characteristics and applications of Twitter content. First, one analyzes the characteristics of Twitter content and then provides basic information for the applications of the content [4, 9, 19]. Cha *et al.* [4] analyzed the characteristics of Twitter content: the number of followers, the number of retweets, and the number of mentions. They analyzed these characteristics to measure a user's influence and found that these characteristics

imply different types of influence. They found that the number of followers, the number of mentions, and the number of retweets have a different user influence on the analysis of the Twitter content. The number of followers implies the size of the user's audience. This means that if a user has a large number of followers, he/she has a large audience who received his/her contents. The number of retweets shows the content value about continued interest. It also shows the content's worth with respect to being shared. The number of mentions shows that a user is related to other conversations, so it shows the content's advertisement value. Kwak *et al.* [9] researched the spreadability of Twitter and found that Twitter is not only an SNS but also a news media. They found that the number of followers represents the user's popularity, and the number of retweets is a major measurement parameter of the user's influence. Further, they found that Twitter trends match the other news media's trends by about 85%. Therefore, they concluded that Twitter is an SNS and can be considered a news media as well. Teevan *et al.* [19] studied the differences in web search and Twitter search and specified the characteristics of Twitter search. They found that a Twitter user performs a search for obtaining "timely information," "social information," and "topical information." They also found that Twitter search has its own characteristics. For example, queries of Twitter search are shorter than those of web search, but have longer words. Further, Twitter search allows the use of site-specific grammar such as the use of "@" and "#." In our study, we applied three factors found in previous research for evaluating the proposed method for the measurement of the influence of Twitter content. These factors are the number of followers, the number of retweets, and the posting time.

The second field is the application and analysis of Twitter content. There is trending extraction research [3, 15, 17, 21] and sentimental analysis [1, 2, 6, 14, 18, 20, 22]. For example, in an analysis field study, Benhardus and Kalita [3] investigated the trend detection in Twitter. They applied frequency, term frequency–inverse document frequency (tf–idf), and normalized term frequency for detecting trends in Twitter content. They found that natural language processing tools are suitable for analyzing Twitter content. Phelan *et al.* [15] proposed a system for the recommendation of articles by using Twitter and the tf–idf score. They also showed that their prototype system (Buzzer) provides the recommendation result for a user and experimentally proved the suitability of their method. Song and Kim [17] proposed a Twitter trend mining system to provide real-time trends. Their system mines social trends and generates a content-based network; their case study was on the 2012 Korean presidential election. Weng *et al.* [21] proposed a method for identifying an influential user of Twitter on the basis of the PageRank algorithm. In the field of sentimental analysis, many researchers have proposed methods to analyze sentiments from Twitter content [1, 2, 6, 14, 18, 20, 22]. Although these previously proposed methods are excellent, their methods are domain-specific and are difficult to apply to every situation. This implies that sentiments are difficult to predict and analyze. However, previous research does show that Twitter can have an infinite number of applications.

In this paper, we propose a method for analyzing trends, particularly popularity, by using Twitter characteristics, and show that the basic methods used in previous studies are useful for Twitter content analysis (TCA). Further, we determine which characteristics are useful for Twitter content analysis.

3. Framework and Method for Twitter Content Analysis

In this section, we introduce our previous research and framework of this study. In Sec. 3.1 we introduce our previous research for measuring Twitter contents influence briefly. We describe our framework in Sec. 3.2 and the analyzing method based on the Twitter content influence in Sec. 3.3.

3.1. *Method for measuring content influence*

The Cambridge Dictionary defines influence as "the power to have an effect on people or things, or a person or thing that is able to do this." There are a large number of theories for measuring influence in the sociology field. However, a method for measuring influence in SNS would be different from previous theories in sociology [4]. Therefore, in our previous research [10], we assumed that a single Twitter content has its own influence on users, that those influences are different from each other by their own characteristics, and that influential content contains meaningful information for users who are exposed to the Twitter content. Therefore, the previous study defines the influence of content as follows: "Content influence is a value that measures to what degree a piece of content contains meaningful information for users" [10]. The study proposes an equation that consists of three characteristics of Twitter: the number of followers, the number of retweets, and the posting time. The number of followers represents spreadability. When a user posts some content, it is primarily spread to the user's followers. The number of retweets denotes shareability because the retweet mechanism is a method of sharing content in Twitter. Further, the value of content can be measured by the number of shares, because if the content is valuable, it is shared with others. Therefore, the number of retweets also denotes the value of the content. The posting time indicates how current the information is. Twitter is sensitive to up-to-date information, so this factor is used for measuring the content's influence [7, 9]. The equation of measuring content influence using the abovementioned three factors is as follows [10]:

$$I(C_i) = \alpha \log(RT_i + 1) + \beta \log(F_i + 1) + \gamma \log\left(\frac{k}{NT - WT_i}\right), \quad \alpha + \beta + \gamma = 1, \quad (1)$$

where C_i represents the ith content and I denotes the influence of C_i. RT_i represents the number of retweets of C_i, so $\log(RT_i + 1)$ is the shareability of C_i. It takes a logarithm function to normalize, and adds 1 to prevent an output that is not negative infinity. F_i represents the number of followers of the author of C_i so $\log(F_i + 1)$ denotes the shareability. It takes a logarithm function for the same reason as that mentioned

earlier. Further, $\log(\frac{k}{NT-WT_i})$ represents how current the information is. WT_i represents the posting time of C_i, and NT denotes the current time now, which is the start time of the analysis. The k-value represents the criteria of determining how current the posting time is. If $(NT - WT_i)$ is smaller than k-hour, then $\log(\frac{k}{NT-WT_i})$ will be a positive value. In contrast, if $(NT - WT_i)$ is greater than k-hour, then $\log(\frac{k}{NT-WT_i})$ will be a negative value. This parameter also takes a logarithm function to normalize. The terms α, β, and γ are mediators used for adjusting the power of each factor. In the Twitter influence equation, it appears that there is a linear relationship between followers and retweets. However, in experiments from a previous research [10], there was no correlation observed between a follower and retweet, using the Pearson correlation coefficient. In addition, for the experimental data in this paper, only 0.06% of the contents are retweeted. Therefore, we assume that retweeted contents have more influence than others, and the influence equation can be used to give more weight to retweeted contents for content evaluation. We evaluated this method using Twitter data and found that it can accurately measure the influence of a single Twitter post [10].

3.2. *Framework*

In this subsection, we describe a framework for content analysis. Figure 1 shows the proposed framework for content analysis.

We crawled content and relation data using Twitter4J API [24], and part of the data is provided by the Daumsoft Company. The crawled data consisted of content information (i.e. authorID, number of retweets, posting time, and content) and user information (i.e. authorID, follower id, and followee id). First, we crawled the content information and saved it as a JavaScript Object Notation (JSON) [25] data type

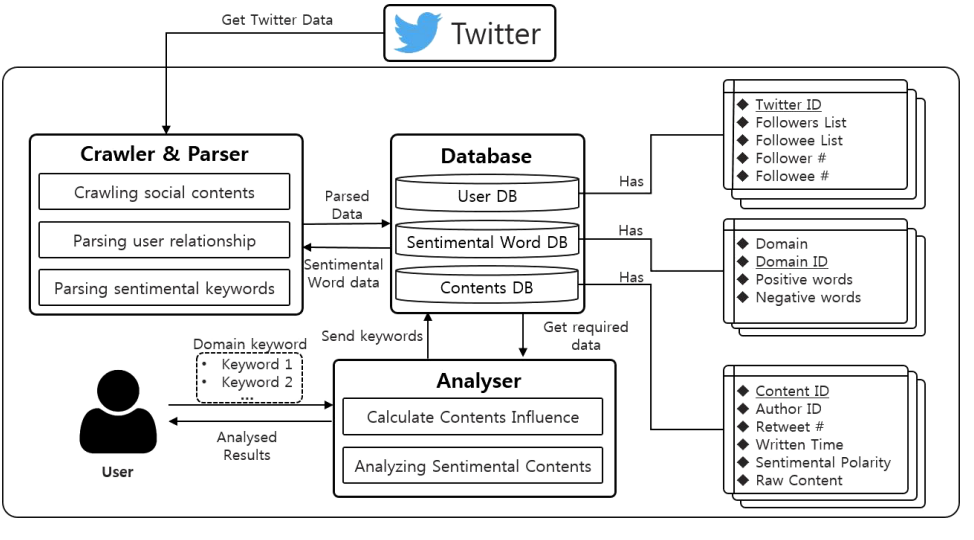

Fig. 1. Framework for Twitter analysis.

```
{
"documentId" : "224855344663371000",
"content" : "싸이 강남스타일 노래 짱인듯♥",
"writtenTime" : "20120716221746",
"shareCount" : 0,
"authorId" : 4384302**,
"displayAuthorId" : "yuso07**"
}
```


Fig. 2. Example of a crawled content file.


```
21122393        190     1824
21174996        346     66
21209183        304     436
21309740        617     616
21366823        554     1498108
```


Fig. 3. Example of crawled follower and follower number file.


```
365516640       139168918,344801362,158487331,136007061,155927976,224223563
365516660       11348282,73992972,51658431,15907720
365516680       130122646,136007061,14872237,246225682,51658431
```


Fig. 4. Example of relationship file.

(see Fig. 2). The crawled JSON file consisted of documentId, content, writtenTime (i.e. posting time), sharedCount (i.e. retweet count), authorID, and displayAuthorId. After content crawling, we crawled additional information about the users. We crawled the number of followers, the number of followees, and the list of followees for each user. This additional information was saved in the simple text format. Figure 3 shows the number of followers and the number of followees for the users. The first term in a line denotes the user's id; the second, the number of followees; and the third, the number of followers. Figure 4 shows the user–followee relationship. The first term in a line denotes the user's id, and second array denotes user's followee list.

After crawling, the crawled data are processed in a parser and saved in a database. Figure 5 shows the parsing and saving processes. The parser creates a network of user relationships and finds sentimental keywords that are defined in the sentimental word database (details of sentimental keywords are provided in Sec. 4.1). While creating this network, it finds a user's followee in the crawled user's followee list and saves the list of followees. Further, it finds the inverse followee list to figure out the followers of the user, as followers can be found by an inverse relation of the followee list. After follower and followee lists processing, the parser finds the number of followers and that of followees in the crawled user's followee and follower number data.

During the content database creation process, the parser extracts information about documentId, authorId, sharedCount (i.e. retweet number), posting time

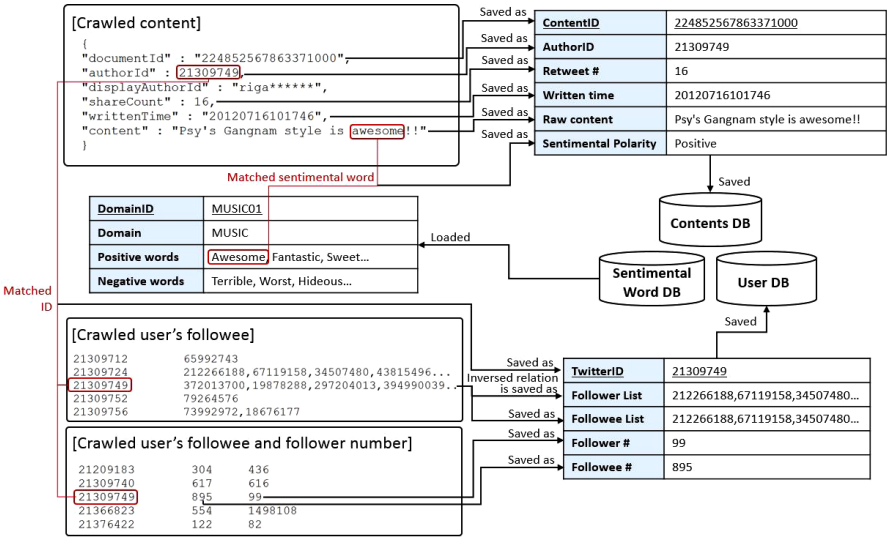

Fig. 5. Example of parsing and saving processes in parser.

(i.e. writtenTime), content, and sentiment polarity. For obtaining sentiment polarity, the parser uses the sentiment database and finds matching words related to sentimentality. If there are any matching positive or negative words, it sets the polarity of the content. Examples of positive words are "good," "excited," "sweet," and "mellow." Some examples of negative words are "terrible," "worst," "hideous," and "indistinctive".

After parsing, the parsed data are saved in a database that consists of three databases: user, sentiment, and content. The user database saves the relationship between user data and has values such as user id (i.e. key value), follower user's id, followee user's id, followee number, and follower number. The sentiment word database describes positive and negative words for a specific domain and has values that are the domain of the sentiment keywords, id of the domain (i.e. key value), positive words, and negative words. Section 4.1 describes the sentiment database in detail. The last database is the content database. It describes the details of a single post and consists of the content id (i.e. key value), authorID, the number of retweets, posting time (i.e. writtenTime), sentiment polarity, and raw content. The analyzer takes the keywords provided by a user and then extracts the data from the database. It then calculates the content influence and analyzes the sentiment value of the keywords. The analyzed result is sent to the user. Section 3.3 describes this method in detail.

3.3. *Method for Twitter content analysis*

To analyze Twitter content, we propose a new approach based on the method of measuring Twitter content influence [10]. Frequency is a simple factor for measuring keyword popularity. Google Trends [8] uses the web search keyword frequency to

extract trends of a specific period. Further, inverse document frequency (idf) is a factor to measure term specificity in a document set. It is widely used in the field of numerical statistics [11, 16]. We use these factors along with the Twitter content influence for the analysis of Twitter content.

In Eq. (1), the k-value denotes a time period factor, and it gives bigger influence score in a k-hour [10]. Therefore, we assume that if we use the accumulated data obtained at the start of analysis, then the popularity of keywords is more accurate than if we only consider a specific period because if a keyword comes up frequently in the past but not in the recent times, we can conclude that the keyword has lost its popularity. In the influence equation, if the content is created out of the k-hour, then the time factor gives a smaller score than that for content created within the k-hour [10]. This characteristic reflects the changes in the content time influence. Therefore, we use this characteristic and propose a method that uses the accumulated content influence with frequency. Equation (2) describes this method:

$$\mathrm{TCA}(k_i, \mathrm{Start}, \mathrm{End}) = \sum_{\mathrm{Start}}^{\mathrm{End}} \left\{ \log \frac{|C|}{|c \in C, \mathrm{Contain}(k_i \in c)|} \times I(\mathrm{Contain}(k_i, c)) \right\}. \quad (2)$$

This equation provides a summation of the influences with idf of the ith keyword between the start and end of the analysis period. Parameters "Start" and "End" denote the start and end times of the analysis period. Therefore, the sigma operator represents a summation of influences that contain k_i with idf from the start time to the end time. C represents the content dataset, and $\mathrm{Contain}(k_i, c)$ denotes the content that includes the ith keyword. Therefore $\log \frac{|C|}{|c \in C, \mathrm{Contain}(k_i \in c)|}$ represents the idf of k_i. $I(c_i)$ indicates the content influence c_i. Finally, the equation describes the popularity of the keyword within the entire analysis period.

4. Experiments and Results

We crawled Korean Twitter content and set up the experimental environment introduced in Sec. 3.1. Here, we describe the experimental dataset in Sec. 4.1 and present the experimental results in Sec. 4.2.

4.1. *Experimental data*

We crawled the contents, user information, and relations of Korean Twitter posts posted in the site between 1 July 2012 and 31 July 2012 (as mentioned in Sec. 3.1, part of the data is provided by Daumsoft). The crawled data are saved in the JSON format first. The size of the content data is about 93 GB, and that of the user and relationship is 12 GB. For the experiments, we chose the music and movie domains because their chart changes every day and every week. Moreover, many reliable online music sites provide music charts based on their own policies, and the popularity of a movie can be measured by ticket sales. Therefore, we can compare the popularities evaluated by the proposed method and the music charts by using online music sites

and movie ticket sales. In order to do it, we collected nine keywords from a popular Korean music website called "Melon" [12], and six keywords from the "Korean Film Council" [13]. The Melon chart is reasonable for measuring popularity because Melon measures its daily music chart by using two factors: how many streams were played and how many mp3 files were downloaded. Melon gives different weights to streaming (40%) and mp3 download (60%). In addition, the Korean Film Council is a government organization of the Republic of Korea, and they provide movie charts that are based on ticket sales. Further, we collected posts on music and movie domains that were released in July 2012, and extracted 71,837 and 239,635 posts related to the considered music and movie domains. Tables 1 and 2 list the Korean words considered for the extraction.

Table 1. Music words considered in the experiment.

Keyword #	Artist	Album title	Words (Korean and English)	Note
1	SISTAR	Loving U	LOVING U ‖ LOVINGU ‖ 러빙유	Not case sensitive
2	Huh Gack	One person	(허각) && (한사람‖한 사람)	—
3	PSY	Gangnam style	(강남 스타일)‖강남스타일‖ GANGNAM STYLE ‖ GANGNAMSTYLE	Not case sensitive
4	T-ara	Day by day	DAYBYDAY ‖ DAY BY DAY ‖데이바이데이‖데이 바이 데이	Not case sensitive
5	F(x)	Electric Shock	(ELECTRIC SHOCK) ‖ (ELECTRICSHOCK) ‖ (일렉트릭쇼크)‖(일렉트릭 쇼크)‖ 전기충격	Not case sensitive
6	VerbalJint	Pretty enough	충분히예뻐‖충분히 예뻐	—
7	Wonder girls	Like this	(원더걸스 ‖ WONDERGIRLS ‖ WONDER GIRLS ‖ 원더걸즈) && (LIKE && THIS)	Not case sensitive
8	Lee Hyun	Heart broken	가슴이 && (시린게‖시린 게)	—
9	2NE1	I love you	(2NE1 ‖ 투애니원) && (I LOVE YOU ‖ ILOVEYOU)	Not case sensitive

Table 2. Movie words considered in the experiment.

Keyword #	Movie name	Words (Korean and English)	Note
1	Deranged	(연가시) && (영화 ‖ movie ‖ 무대‖시사회‖김명민)	Not case sensitive
2	Madagascar 3	마다가스카	—
3	A Letter to Momo	(모모 && 다락방 && 요괴) ‖ A Letter to Momo)	Not case sensitive
4	The Amazing Spider-Man	(어메이징 && 스파이더맨)‖스파이더맨 ‖ spiderman ‖ spider man)	Not case sensitive
5	All About My Wife	(내 아내의 모든 것‖내아내의모든것‖내 아내의 모든것)	—
6	Ice Age 4	(아이스 && 에이지 && 4)	—

Table 3. Music keywords rankings.

Album title #	1st Week	2nd Week	3rd Week	4th Week
Loving U	1	2	3	3
One person	6	8	10	15
Gangnam style	—	—	1	1
Day by day	4	3	4	5
Electric Shock	5	6	11	17
Pretty enough	10	10	14	16
Like this	8	11	17	21
Heart broken	7	5	7	8
I love you	9	1	2	2

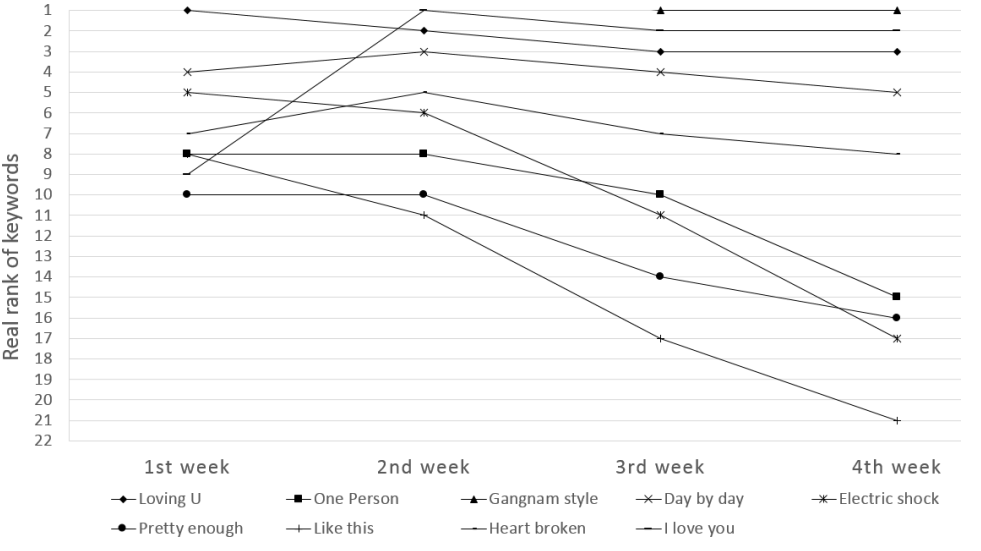

Fig. 6. Variations of music keywords ranks.

Table 3 and Fig. 6 describe the rankings of these nine albums, which are provided by Melon [12]. We chose keywords that had different weekly trends. For example, "Loving U," "Day by day," and "I love you" maintained a high rank every week, and "One person", "Electric Shock," "Pretty enough," and "Like this" had a high rank at the beginning of the month but fell to a low rank later in the month. Lastly, "Gangnam style" appeared suddenly in the third week and maintained a high rank. Because this keyword was published on 12 July 2012, there were no rank values for it during the first and second weeks.

Table 4 and Fig. 7 describe the rankings of the movie domain, which are provided by the Korean Film Council [13]. We chose five keywords (i.e. "Deranged," "Madagascar 3," "A Letter to Momo," "The Amazing Spider-Man," and "All About My Wife") that had high rankings at the beginning of the month but fell during the

Table 4. Movie keywords rankings.

Movie title	1st Week	2nd Week	3rd Week	4th Week
Deranged	2	1	2	4
Madagascar 3	6	8	13	24
A Letter to Momo	5	5	9	17
The Amazing Spider-Man	1	2	3	7
Ice Age 4	32	26	55	3

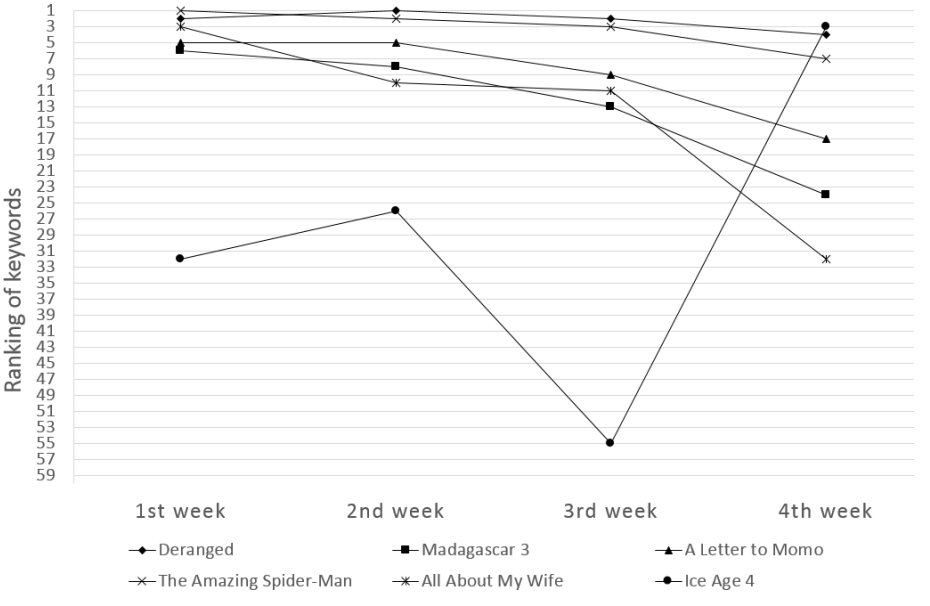

Fig. 7. Variations of movie keywords ranks.

month. "Ice Age" was ranked low at the beginning of the month, but suddenly rose in rank at the end of the month.

For the sentiment analysis with sentiment polarity, positive and negative words are needed because if a post has keywords with positive sentiment words, it is considered to contain positive sentiment, and if the post has keywords with negative words, it is considered to contain negative sentiment. Therefore, for measuring sentiment polarity, we need to define positive and negative words related to music and movie domains. For this, we collected Korean positive and negative words related to music and movie domains by using WordNet [5]. The total number of words considered in the music domain was 402; there were 220 positive words and 182 negative words. In addition, the total number of words considered in the movie domain was 436; there were 225 positive words and 211 negative words. Tables 5 and 6 list these words in detail, and the measurement method is described in Sec. 5.1.1.

Table 5. Positive and negative Korean words related to music domain.

Positive Korean words (220)	Negative Korean words (182)
가슴 벅차다, 간지나다, 감동받다, 강렬하다, 갖고 싶다, 고급스럽다, 고맙다, 고져스하다, 고풍스럽다, 곱다, 괜찮다, 굉장하다, 구입하고싶다, 귀엽다, 귀중하다, 귀하다, 근사하다, 기대, 기발하다, 기분 전환, 기분 좋다, 기쁘다, 깜찍하다, 깨끗하다, 꼼꼼하다, 끌리다, 내추럴, 노이즈 없다, 다양하다, 대단하다, 두근두근하다, 뒤떨어지지않다, 또렷하다, 뛰어나다, 러블리하다, 럭셔리, 마음 들다, 만족스럽다, 맑다, 매끈하다, 매력 있다, 멋있다, 모던하다, 무난하다, 문제 없다, 믿다, 바람직하다, 반갑다, 반짝반짝하다, 보람 있다, 부드럽다, 분위기 있다, 분위기 좋다, 빛나다, 빼어나다, 뿌듯하다, 사고싶다, 사랑스럽다, 살맛 나다, 상당하다, 새롭다, 생기 있다, 선호, 설레다, 성공, 세련되다, 섹시하다, 소중하다, 속시원하다, 입에감기다, 스타일리시하다, 시간때우기 좋다, 시원하다, 신나다, 신뢰, 싱그럽다, 아기자기하다, 아담하다, 아름답다, 앞서나가다, 어색하지않다, 어울리다, 엄청나다, 업되다, 엑설런트, 여유롭다, 예쁘다, 오래 가다, 올바르다, 옳다, 완벽, 우수하다, 우아하다, 우월하다, 월등하다, 위대하다, 유니크하다, 유리하다, 유명하다, 유쾌하다, 유행, 은은하다, 음질 좋다, 이롭다, 인기많다, 인기 있다, 인정, 자연스럽다, 잘나가다, 잘나오다, 잘되다, 잘뜨다, 잘만들다, 재미있다, 적당하다, 적합하다, 전망 밝다, 정성스럽다, 제대로 되다, 조마조마하다, 조화롭다, 주목 받다, 즐겁다, 즐길수있다, 진화, 질리지않다, 짜릿하다, 착하다, 참신하다, 천진난만하다, 추천, 출중하다, 충분하다, 친근하다, 친밀하다, 친절하다, 탁월하다, 투명하다, 트렌디, 특별하다, 특이하다, 특화, 패셔너블하다, 퍼펙트, 평가 좋다, 포스 있다, 폭 넓다, 품질 좋다, 풍부하다, 풍성하다, 핫하다, 행복하다, 향상, 현명하다, 호평 받다, 화려하다, 화사하다, 환하다, 활기차다, 활발하다, 황홀하다, 후련하다, 훈훈하다, 훌륭하다, 흐뭇하다, 흥겹다, 흥분, 희귀하다, 다행, 대박, 대세, 도시적, 안심, 안정적, 업그레이드, 열심, 열정적, 완소, 웰메이드, 이국적, 인상적, 자신감, 정상, 정통, 진리, 짱, 최고, 최신, 최초, 파격적, 합격점, 현대적, 환상적, 고화질, 기대이상, 호감, 잔잔하다, 긍정적, 고음질, 강하다, 빠져있다, 소름 돋다, 울리다, 목소리 좋다, 흥얼거리다, 흥얼대다, 평화, 슬프다, 음악성 좋다, 리듬 좋다, 소리 좋다, 좋다, 사랑하다, 흥하다, 쩔다, 쩐다, 자주 듣다, 감수성 터지다, 달달, 명곡	거지같다, 걱정스럽다, 격분, 고발, 고통스럽다, 관심 없다, 괴롭다, 구리다, 그지 같다, 기분 나쁘다, 꽂히지않다, 낚이다, 낯설다, 너무길다, 너무짧다, 너무하다, 느껴지지않다, 단조롭다, 답답하다, 더럽다, 둔하다, 뒤떨어지다, 디스, 따라하다, 따분하다, 뜨지않다, 루즈하다, 맞지않다, 모방, 무한반복되다, 문제 있다, 미치다, 믿을수없다, 밋밋하다, 반응없다, 듣기 불편하다, 복잡하다, 부실하다, 부자연스럽다, 부족하다, 분개, 분노, 분하다, 불가능하다, 불만 많다, 불안정하다, 불안하다, 불친절하다, 불쾌하다, 비난 받다, 빈티나다, 사칭, 산만하다, 성차지않다, 성가시다, 성질나다, 소리 너무작다, 소리 너무크다, 소리 맞지않다, 소리 이상하다, 소송, 소음, 속상, 속이다, 시끄럽다, 식상, 실망, 실수, 실패, 싫어하다, 심각하다, 심하다, 들을일없다, 아깝다, 아쉽다, 안쓰럽다, 약하다, 어렵다, 어리둥절하다, 어색하다, 어울리지않다, 어이없다, 어지럽다, 얼떨떨하다, 열받다, 예쁘지않다, 올드하다, 요상하다, 욕먹다, 욕하다, 우려, 울리지않다, 원망, 위태롭다, 위험하다, 유치, 유행 지나다, 의심, 이상하게 나오다, 이상하다, 익숙하지않다, 인식 되지않다, 재미없다, 적응하기어렵다, 정 가지않다, 정신 없다, 제재받다, 조악하다, 조잡하다, 주저, 지루하다, 지치다, 집단 소송, 징계, 짜증 나다, 찝찝하다, 초라하다, 촌스럽다, 콘텐츠싸움 번지다, 탁하다, 탈퇴, 토 나오다, 투박하다, 특색 없다, 필요 없다, 한계많다, 한계 있다, 한심하다, 해롭다, 허무하다, 허전하다, 허탈하다, 혹평 받다, 화나다, 황당하다, 효과없어지다, 후회, 흔들리다, 흠집 나다, 힘들다, 구식, 근심, 꽝, 단점, 멘탈붕괴, 별로, 병맛, 불법, 비호감, 상술, 상업적, 싸구려, 엽기적, 장삿속, 최악, 호구, 무책임하다, 손실률 많다, 위반, 제한, 차단, 침해, 허용하지않다, 끔찍하다, 징그럽다, 듣기싫다, 기대이하, 기대 미치지못하다, 꼬이다, 마음 들지않다, 촌극, 돈지랄, 혐오, 혐오스럽다, 엿 같다, 거슬리다, 부정적, 표절, 못 듣겠다, 좋지 않다, 의혹, 젠장

4.2. *Experimental results*

First, we adjust mediators i.e. α, β, and γ in Eq. (1) that adjust the three influence factors for the optimization of proposed method (see Sec. 3.2). These mediators are shareability, spreadability, and currency of time. These are indicated by α, β, and γ

Table 6. Positive and negative Korean words related to movie domain.

Positive Korean words (225)	Negative Korean words (211)
가능하다,가슴 벅차다,간지나다,감동받다,강렬하다,걱정 없다,견고하다,고급스럽다,고맙다,고져스하다,고풍스럽다,곱다,괜찮다,굉장하다,구입하고싶다,귀엽다,귀중하다,귀하다,근사하다,기대,기발하다,기분 전환,기분 좋다,기쁘다,기상천외하다,깜찍하다,깨끗하다,꼼꼼하다,끌리다,내추럴,다양하다,대단하다,두근두근하다,뒤떨어지지않다,따라잡다,뛰어나다,러블리하다,럭셔리,마음 들다,만족스럽다,맑다,매력 있다,멋있다,모던하다,무난하다,문제 없다,믿다,바람직하다,반갑다,반응 빠르다,반짝반짝하다,보람 있다,부드럽다,부럽다,분위기 있다,분위기 좋다,빛나다,빵터지다,빼어나다,뿌듯하다,사고싶다,사랑스럽다,상당하다,새롭다,색감 좋다,생기 있다,선호,설레다,성공,세련되다,섹시하다,소중하다,속시원하다,스타일리시하다,시간때우기 좋다,시원하다,시청 가능하다,신기하다,신나다,신뢰,싱그럽다,아기자기하다,아담하다,아름답다,앞서나가다,어색하지않다,어울리다,엄청나다,업되다,엑설런트,여유롭다,예쁘다,옳다,완벽,용감하다,우수하다,우아하다,우월하다,월등하다,위대하다,유니크하다,유리하다,유명하다,유용하다,유쾌하다,유행,은은하다,이상 없다,인기 많다,인기 있다,인정,자연스럽다,잘나가다,잘되다,잘뜨다,잘만들다,재미있다,적당하다,적합하다,전망 밝다,정성스럽다,정확하다,제공,제대로 되다,조건 좋다,조마조마하다,조화롭다,주목 받다,즐겁다,즐길수있다,진화,질리지않다,짜릿하다,착하다,참신하다,천진난만하다,추천,출중하다,충분하다,친근하다,친밀하다,친절하다,탁월하다,투명하다,트렌디,특별하다,특이하다,특화,패셔너블하다,퍼펙트,평가 좋다,포스 있다,폭 넓다,품질 좋다,풍부하다,풍성하다,핫하다,행복하다,향상,호평 받다,화려하다,화사하다,환하다,활기차다,활발하다,황홀하다,후련하다,훈훈하다,훌륭하다,흐뭇하다,흥겹다,흥분,희귀하다,다행,대박,대세,도시적,안심,안정적,업그레이드,열심,열정적,완소,웰메이드,이국적,인상적,자신감,정상,정통,진리,짱,최고,최신,최초,파격적,합격점,현대적,환상적,기대이상,호감,잔잔하다,긍정적,강하다,긴장감 있다,뜨겁다,소름 돋다,울리다,감동적이다,공감하다,공감가다,한번 더보다,보고싶다,잘생겼다,매력,흥미진진,흥미,잘만들다,보러가고싶다,관객 많다,연기 잘하다,명품,웃기다,빠져들다,쩐다,반하다	거지같다,거추장스럽다,걱정스럽다,겁먹다,격분,격하다,겹쳐보이다,고민,고발,고통스럽다,관심 없다,광고 싫다,괴롭다,구리다,귀찮다,그지 같다,기분나쁘다,까다롭다,깨끗하지않다,낚이다,날아가다,낯설다,너무길다,너무짧다,너무약하다,너무하다,눈에 들어오지않다,눈 아프다,피곤하다,느껴지지않다,단조롭다,답답하다,더럽다,뒤떨어지다,디스,따라하다,따분하다,뜨지않다,루즈하다,리콜,망설이다,맞지않다,모방,문제 있다,밋밋하다,반응 없다,보기 불편하다,복잡하다,볼수없다,부실하다,부자연스럽다,부족하다,분개,분노,분하다,불가능하다,불리하다,불만 많다,불안정하다,불안하다,불쾌하다,불편하다,비난 받다,빈티나다,산만하다,성 차지않다,성질나다,소리 너무작다,소리 너무크다,소리 맞지않다,소리 이상하다,소음 들리다,속상,속이다,시간 낭비,시끄럽다,식상,실망,실수,실패,싫어하다,심각하다,심하다,아깝다,아쉽다,안쓰럽다,약하다,어렵다,어리둥절하다,어색하다,어울리지않다,어이없다,어지럽다,얼떨떨하다,열받다,예쁘지않다,오락가락하다,올드하다,요상하다,욕먹다,욕하다,우려,우유부단하다,울리지않다,원망,위태롭다,위험하다,유출,유치,유해하다,유행 지나다,음질 나쁘다,의심,이상하게 나오다,이상하다,익숙하지않다,재미없다,적응하기어렵다,정 가지않다,정신 없다,제공 되지않다,제외,제재받다,제한 있다,조악하다,조잡하다,지루하다,지연,지치다,집단 소송,징계,짜증 나다,찝찝하다,초라하다,촌스럽다,촛점 흐리다,충돌 있다,탁하다,토 나오다,투박하다,특색 없다,품질 바뀌지않다,피해 입다,필요 없다,한계 많다,한계 있다,한심하다,해롭다,허무하다,허전하다,허탈하다,현기증 나다,혹평 받다,화나다,확인 되지않다,황당하다,효과 없어지다,후회,휑하다,흔들리다,흠집 나다,힘들다,구식,근심,꽝,단점,말썽,멘탈붕괴,별로,병맛,불량,불법,불통사태,비호감,상술,상업적,상영불가,상영중단,싸구려,최악,호갱님,호구,무책임하다,손실 많다,위반,제한,침해,끔찍하다,징그럽다,보기싫다,악성,기대이하,기대 미치지못하다,마음 들지않다,촌극,돈지랄,섬뜩하다,혐오,혐오스럽다,엿 같다,거슬리다,부정적,뻘짓,기대 안하다,뻔하다,졸리다,막장,연기 못하다,발연기,싫다,별로다

in Eq. (1). We changed the weight of each factor using the experimental dataset described in Sec. 4.1. We performed an optimization experiment using only the music domain experiment dataset. We applied optimized values for different domains to show that the optimized value is valid not only in the music domain but also in different domains (i.e. movies). Optimization results were obtained by a comparison

of the Melon chart (see Sec. 4.1) and the ranking computed by the proposed method. We compared both rankings for each week of July and used an average of four weeks. Spearmans rank correlation coefficient was used for the comparison. This method has a value between 1 and 1. If the absolute value was close to 1, there was a significant linear relationship. However, if the absolute value was close to 0, there was a non-linear relationship between the two datasets. In this method, positive values represent a positive correlation, and negative values denote a negative correlation. Therefore, if the Spearman's rank correlation coefficient result between the Melon chart and the ranking calculated by the proposed method was close to 1, then popularity extracted by the proposed method was well analyzed. Before adjusting all three parameters, the first experiment considered only shareability and spreadability except for time influence. Table 7 and Fig. 8 show the results.

Here, α denotes the shareability mediator and β is the spreadability mediator; γ represents the time mediator, but its value is 0 because this experiment does not consider the currency of time. There is no difference when α is between 0 and 0.3, but when α is between 0.4 and 0.6, the coefficient value increases slightly. Further, the coefficient value decreases when α is close to 1. We can infer that the number of

Table 7. Results of adjusting shareability and spreadability for the proposed method.

	α-value and β-value ($\gamma = 0$)										
Adjusting values	$\alpha = 0$, $\beta = 1$	$\alpha = 0.1$, $\beta = 0.9$	$\alpha = 0.2$, $\beta = 0.8$	$\alpha = 0.3$, $\beta = 0.7$	$\alpha = 0.4$, $\beta = 0.6$	$\alpha = 0.5$, $\beta = 0.5$	$\alpha = 0.6$, $\beta = 0.4$	$\alpha = 0.7$, $\beta = 0.3$	$\alpha = 0.8$, $\beta = 0.2$	$\alpha = 0.9$, $\beta = 0.1$	$\alpha = 1$, $\beta = 0$
Correlation coefficient	0.70	0.70	0.70	0.70	0.71	0.71	0.71	0.69	0.66	0.66	0.66

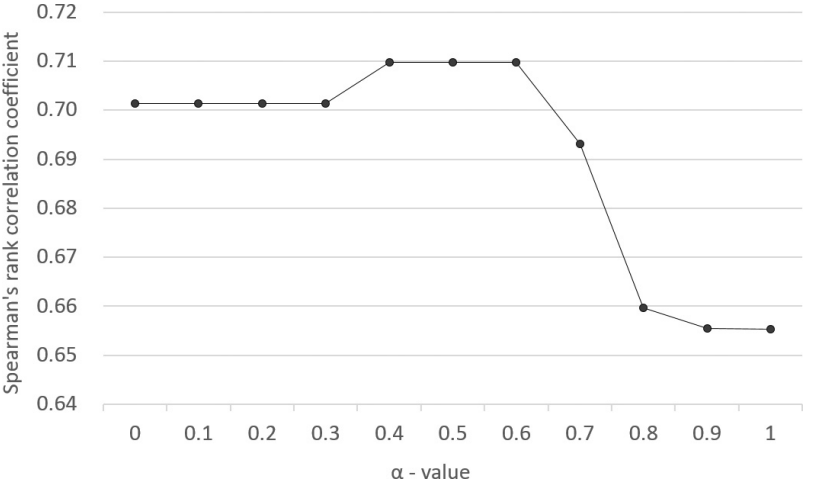

Fig. 8. Results of adjusting shareability and spreadability.

followers (i.e. spreadability) has more influence than the number of retweets (i.e. shareability) on the analysis of popularity. These findings are the same as those of the previous research [10], but the difference is that the previous research considered only retweeted posts, but this research considers not only retweeted posts but also posts that were not retweeted. Moreover, retweeted posts are a minority in this experiment set. Retweeted posts form only 0.06% of the extracted posts. The number of extracted posts is 71,837, and that of the retweeted posts is 4644. Nevertheless, the retweet number impacts the analysis; therefore, we infer that shareability is useful for the analysis of Twitter content, even if it is a minority of contents. However, only considering shareability is worse than considering it along with spreadability.

After optimizing α and β, we performed experiments to adjust the factor γ for the proposed method. Here, γ represents the time currency of the influence equation. We performed the same experiment, but this time we fixed the percentages of α and β and changed the γ values. The fixed percentage of α was 40% and that of β was 60%; these values were based on the results of the previous experiments. Table 8 and Fig. 9 present the results.

Table 8. Results of adjusting shareability, spreadability, and time for the proposed method.

	α-value, β-value, and γ-value										
Adjusting values	$\alpha = 0.4$, $\beta = 0.6$, $\gamma = 0$	$\alpha = 0.36$, $\beta = 0.54$, $\gamma = 0.1$	$\alpha = 0.32$, $\beta = 0.48$, $\gamma = 0.2$	$\alpha = 0.28$, $\beta = 0.42$, $\gamma = 0.3$	$\alpha = 0.24$, $\beta = 0.36$, $\gamma = 0.4$	$\alpha = 0.2$, $\beta = 0.3$, $\gamma = 0.5$	$\alpha = 0.16$, $\beta = 0.24$, $\gamma = 0.6$	$\alpha = 0.12$, $\beta = 0.18$, $\gamma = 0.7$	$\alpha = 0.08$, $\beta = 0.12$, $\gamma = 0.8$	$\alpha = 0.04$, $\beta = 0.06$, $\gamma = 0.9$	$\alpha = 0$, $\beta = 0$, $\gamma = 1$
Correlation coefficient	0.71	0.71	0.71	0.71	0.71	0.71	0.66	0.62	0.61	0.38	0.13

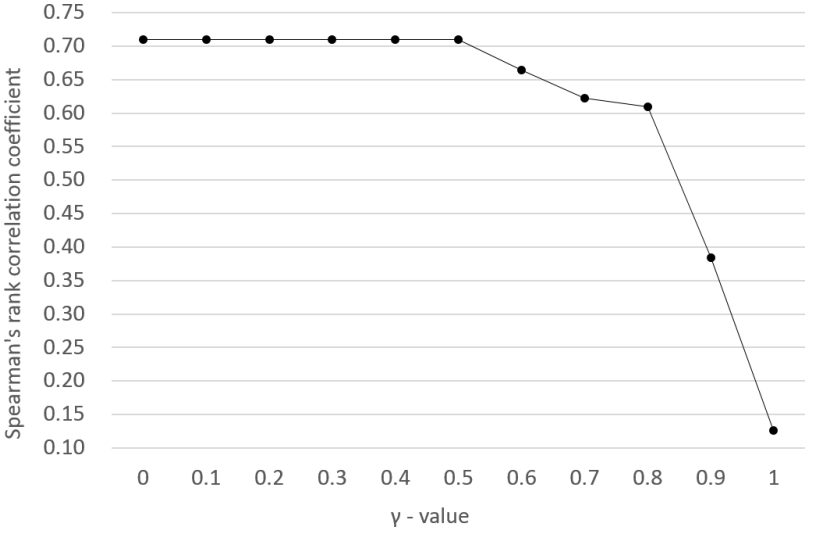

Fig. 9. Results of adjusting γ-value for the proposed method.

The results show that the coefficient value is the best when γ is between 0 and 0.5. However, if we only consider the currency of time, it is not good as considering other factors. These results reveal that not only is time an important factor to consider when analyzing Twitter content, but it is useful when used along with other characteristics. The experiment results show that the coefficient value is the same when γ is between 0 and 5. Therefore, we select the median value as the optimized value of γ weighted by time and other factors (i.e. shareability and spreadability). After these experiments, we obtained the optimization values for the proposed method ($\alpha = 0.32, \beta = 0.48, \gamma = 0.2$).

5. Discussion

In this section, we compare the proposed method with other methods in Sec. 5.1. After the comparison, we discuss characteristics that are useful for the Twitter content analysis in Sec. 5.2.

5.1. *Comparison with other methods*

5.1.1. *Comparison with other methods using total data*

In this sub-subsection, we compare the proposed method with other methods such as the information retrieval method, user influence measurement, sentiment analysis, and frequency measurement. Further, we verify that the proposed method is better than the other methods. To this end, first, we will describe the comparison target methods and then compare the proposed method with the other considered methods. The first method for the comparison is measurement by using frequency. This is a simple and useful method for analyzing Twitter content and searching trends [3, 8]. In this research, we count posts containing keywords. Equation (3) describes this method. It considers the number of posts containing the ith keyword in document set C:

$$\text{frequency}(k_i) = |\text{Contain}(k_i, C)|. \tag{3}$$

The next method considered is the tf–idf [16, 11]. Tf–idf is a numerical statistic that measures how important a word is to a document in a set of documents. It is used for ranking documents in a search engine or a similarity check between different documents. The tf–idf equation is as follows [see Eqs. (4)–(6)]:

$$\text{tf}(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max(f(w, d) : w \in d)}. \tag{4}$$

The factor $\text{tf}(t, d)$ denotes the frequency of term t in document d. Further, idf represents the document frequency containing term t in the document set [see Eq. (5)]:

$$\text{idf}(t, D) = \log \frac{|D|}{|d \in D, t \in d|}. \tag{5}$$

The product of tf and idf is the value of tf–idf:

$$\text{it}-\text{idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D). \tag{6}$$

We crawled 123,639,069 documents posted between 1 July 2012 and 31 July 2012, so we set the total document size "D" as 123,639,069.

For a comparison of the sentiment analysis results, we use the sentiment score described in previous research. Asur and Huberman [2] proposed a sentiment analysis classifier to identify articles that are neutral, positive, or negative. They proposed simple methods for evaluating subjectivity and polarity. The method for evaluating subjectivity is as follows:

$$\text{Subjectivity} = \frac{|\text{Positive and Negative Tweets}|}{|\text{Neutral Tweets}|}. \tag{7}$$

A subjectivity value represents the ratio of contents containing sentiment information. Therefore, the higher the subjectivity value, the larger is the amount of sentiment information in the content set. The method for evaluating the ratio of positive and negative sentiments is as follows:

$$\text{PNratio} = \frac{|\text{Tweets with Positive Sentiment}|}{|\text{Tweets with Negative Sentiment}|}. \tag{8}$$

If the value of PNratio is greater than 1, the majority opinion is positive, but if the value of PNratio is less than 1, the majority opinion is negative. Further, we can obtain the sentiment score by multiplying subjectivity and PNratio:

$$\text{SentimentalScore}(k_i) = \text{Subjectivity}(k_i) \times \text{PNratio}(k_i). \tag{9}$$

We also compared the proposed method with the users influence. For comparison, we modified the original PageRank [23] method on the basis of the follower–followee relationship, substituting the followee for the out-degree and the follower for the in-degree. The initial value of the modified method is the number of followers and the d-value is 0.85 as in the previous research [23]. The d-value represents the residual probability and is usually set to 0.85. The number of authors is 38,968, and the number of connections is 70,225,067. Therefore, we calculate PageRank with 38,968 nodes and 70,225,067 edges. The modified PageRank is as follows:

$$\text{PageRank}(u_i) = \frac{(1-d)}{\text{Total Number Of Users}} + d\left(\sum_{j=1}^{\text{follower Of } u_i} \frac{\text{PageRank}(u_j)}{\text{followee Number Of } u_i}\right). \tag{10}$$

The proposed method is compared with the other methods by using Spearman's rank correlation coefficient. Table 9 and Fig. 10 show the results for the music domain, Table 10 and Fig. 11 show the results for the movie domain.

Table 9. Results of comparison of the proposed method with other methods in music domain.

		Frequency	tf–idf	SentiScore	PageRank	Proposed $\alpha = 0.32$, $\beta = 0.48$, and $\gamma = 0.2$
Spearman's rank correlation coefficient	1st Week	0.59	0.59	0.12	0.61	0.61
	2nd Week	0.58	0.65	0.12	0.58	0.68
	3rd Week	0.67	0.67	0.07	0.60	0.77
	4th Week	0.60	0.68	−0.10	0.48	0.78
	Average	0.61	0.65	0.08	0.57	0.65

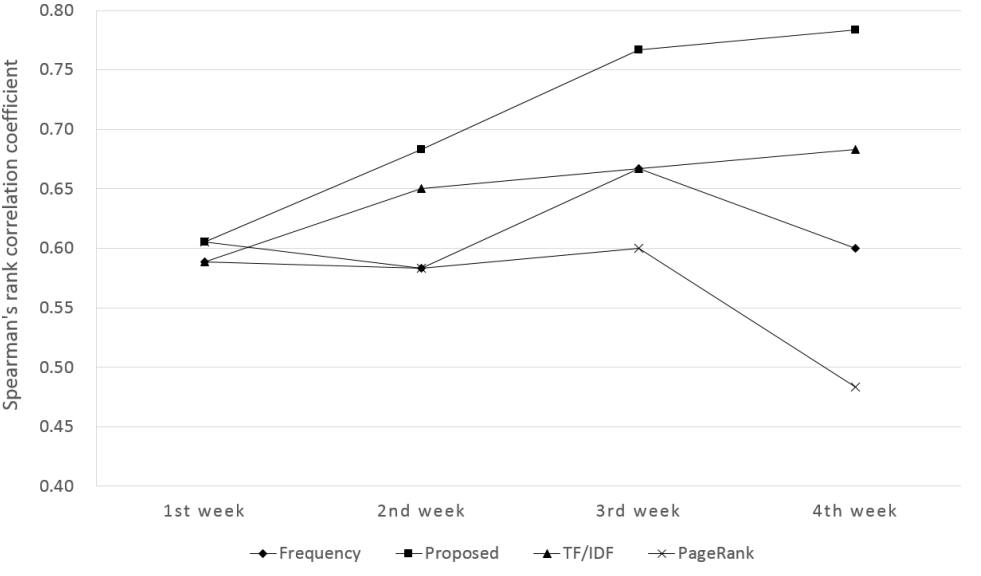

Fig. 10. Results of comparison of the proposed method with other methods in the music domain.

In the experimental data for the music domain, except for the sentiment score, the results show that the proposed method and PageRank have the best coefficient values during the first week of July (see Fig. 10); the other values are similar. Except for the sentiment score, the results show that the proposed method and PageRank have the best coefficient values on 1 July; the other values are similar. However,

Table 10. Results of comparison of the proposed method with other methods in movie domain.

		Frequency	tf–idf	SentiScore	PageRank	Proposed $\alpha = 0.32$, $\beta = 0.48$, and $\gamma = 0.2$
Spearman's rank correlation coefficient	1st Week	0.54	0.54	−0.26	0.60	0.83
	2nd Week	0.60	0.60	−0.71	0.60	0.60
	3rd Week	0.67	0.66	−0.03	0.66	0.66
	4th Week	0.71	0.66	−0.20	0.77	0.77
	Average	0.63	0.61	−0.30	0.66	0.71

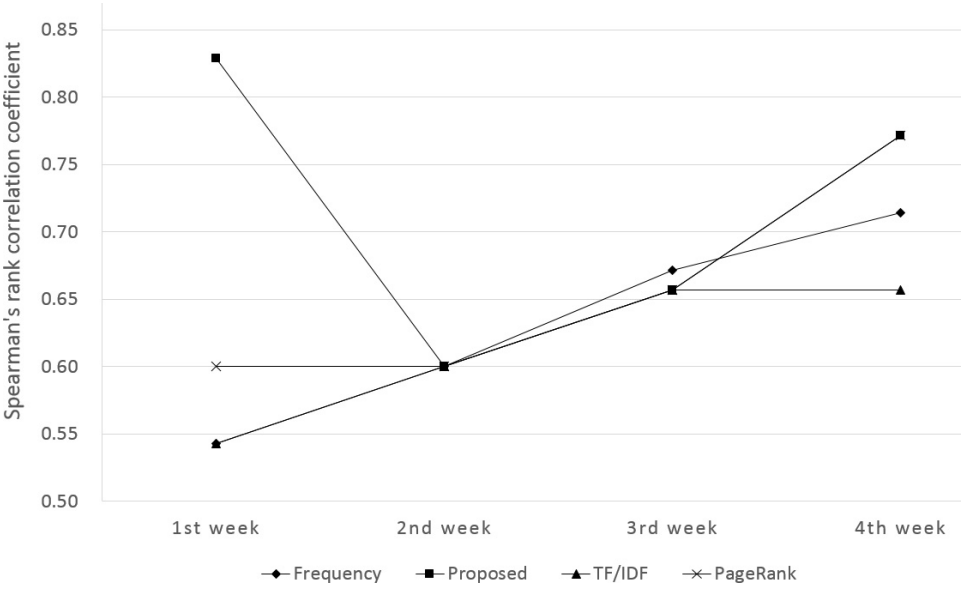

Fig. 11. Results of comparison of the proposed method with other methods in the movie domain.

every method is included in the reasonable range, because the absolute value of the correlation coefficient is between 0.3 and 0.7 and thus shows a moderate correlation relationship. In the second week, the proposed method has the highest value, but all methods are within a reasonable area. In the third week, all methods have reasonable values, but these values decrease dramatically, except for the proposed method. Note that the Melon music chart collects music data from Monday to Sunday and a new chart is posted every Monday. "Gangnam style" was published on Sunday, 15 July 2012, and it suddenly got the first place in the chart. Therefore, all methods have little information about "Gangnam style." However, in spite of the sudden appearance of "Gangnam style," the result of the proposed method was higher than that in the previous week. Further, its value was more accurate, because the absolute value of the correlation coefficient value was greater than 0.7; this implied a strong correlation. In the last week, the coefficient values of PageRank and frequency measurement decreased, and those of the proposed method and the tf–idf value increased.

The movie data results (see Fig. 11) also show that the proposed method has the best coefficient value during the first week of July. The values for other methods are similar and are included in a reasonable range except for the sentimental score. During the second week, all methods have the same coefficient value, and all coefficient values are within a reasonable area. During the third week, the frequency has a higher value. However, there is a very small gap in the proposed method, and the correlation value of the proposed method is within a reasonable area. Further, during

the last week of July, the proposed method and PageRank with higher values according to the data are accumulated. Note that "Ice Age 4" was released on 25 July, and this movie was suddenly ranked third on the chart. For this sudden appearance of "Ice Age 4," the result of the proposed method maintains a strong correlation coefficient. The proposed method is more accurate than the previous methods for the last week of July. This showed us that the proposed method is better than the others when the data are cumulatively collected. Because the proposed method reflects time influence with the criterion of the currency of time, this characteristic had a positive effect on the analysis result when the data were cumulated.

In both experimental datasets, the results of using the sentiment score show that the coefficient value is negative. This indicates that the use of only the sentiment score, considering positive and negative sentiments, is not effective in analyzing popularity. Even ranking by the frequency of negative posts is more correct. To determine why the sentiment score is not a reasonable parameter for the popularity analysis, we used the music domain dataset. Table 11 and Fig. 12 show the numbers of positive and negative posts in music domain. This shows that if some keywords have a large number of positive posts, the others have a large number of negative posts, such as "Loving U" and "Gangnam style." Moreover, even "One person" is not good according to the Melon chart (see Sec. 4.1), but its sentiment score value is better than the others (its rank as measured by the sentiment score is the first or the second). Because "One person" has a firm fan-following and is not often mentioned as popular music, its negative posts are fewer than those of the others; thus, it interrupts the popularity analysis. On the basis of the results of this experiment, we know that sentiment information divided only into positive and negative sentiments is not helpful for the analysis and the ratio of keywords is not always correct.

We calculated the coefficient averages of the proposed method and the other methods (see Figs. 13 and 14). First of all, methods using sentiment information have unacceptable results, and the other methods have reasonable results with a moderate correlation. Further, the proposed method increases the coefficient value more than the other methods with a strong correlation value (greater than 0.7).

Table 11. Numbers of positive and negative posts about keywords in the music domain.

Week	Loving U	One person	Gangnam style	Day by day	Electric Shock	Pretty enough	Like this	Heart broken	I love you
				Positive					
1st Week	2043	120	1	803	454	399	60	44	641
2nd Week	1737	66	238	547	327	287	31	25	335
3rd Week	519	27	1000	252	203	100	8	23	112
4th Week	370	31	1529	208	152	147	14	20	108
				Negative					
1st Week	93	7	0	60	45	21	3	4	23
2nd Week	64	2	18	42	19	14	3	7	18
3rd Week	92	1	95	25	21	8	2	3	10
4th Week	189	5	442	53	42	21	5	5	21

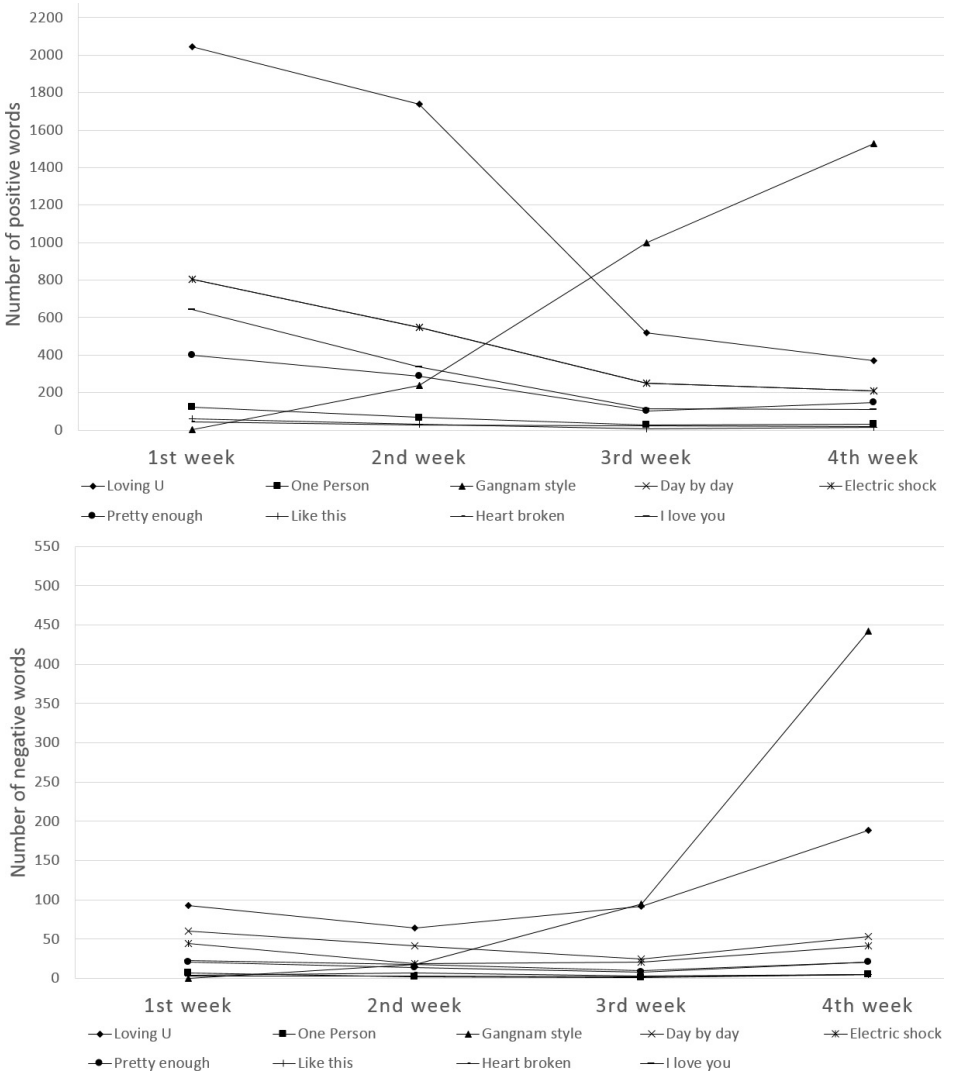

Fig. 12. Graphs of the numbers of positive and negative posts about keywords.

5.1.2. *Comparison with other methods using different data sizes*

One problem in studying Twitter content analysis is that a large amount of data is needed to analyze Twitter contents. However, not every researcher can collect a large amount of data, so a Twitter content analysis method is required to analyze Twitter contents using small data sizes. In this sub-subsection, we performed experiments for the proposed method using different content sizes for which the proposed method is reasonable even for a small amount of data. For the experiment, we assigned an ID number between 0 and 69999 for 70,000 pieces of content in the music domain data

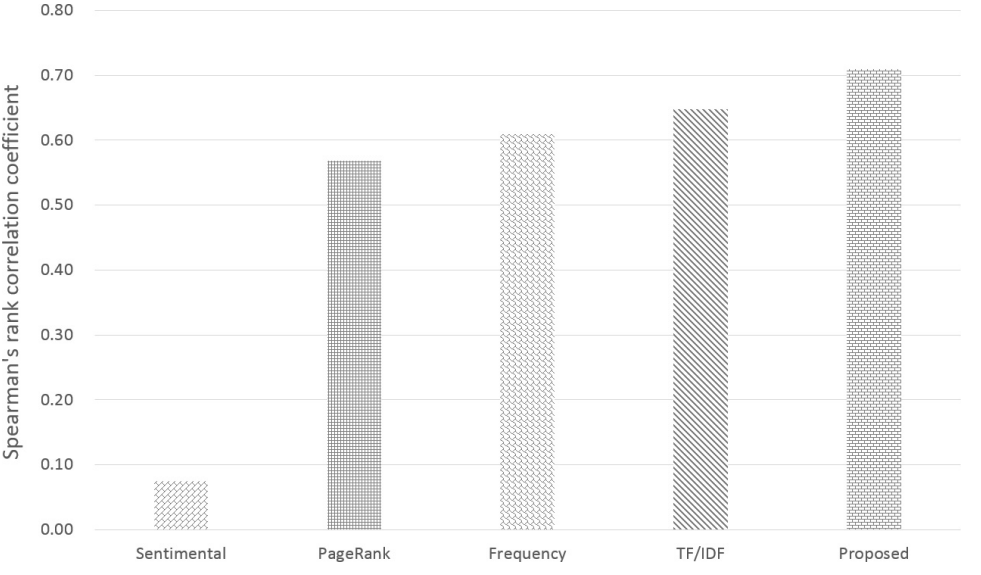

Fig. 13. Average comparison of the proposed method and other methods in the music domain.

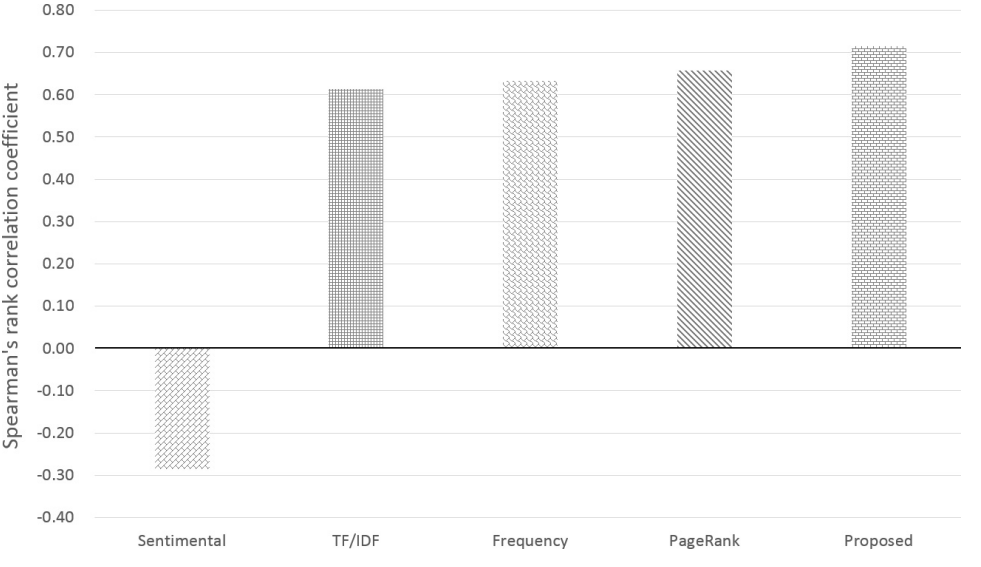

Fig. 14. Average comparison of the proposed method and other methods in the movie domain.

set. Each content number is randomly assigned and is not duplicated. After assigning, we divided the data size by using the assigned content number. We classified data sizes as 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 10,000, 20,000, 30,000, 40,000, 50,000, 60,000, and 70,000. In addition, experimental data consisted

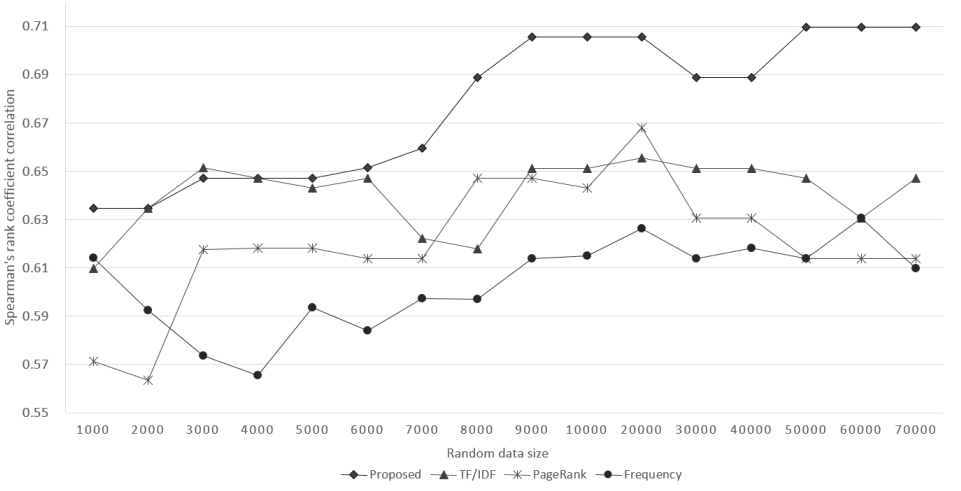

Fig. 15. Results of random sampling.

of contents that were assigned content numbers based on data size. For example, if the data size is 10,000, this dataset consists of contents that have assigned content numbers 0–9999. We performed the same experiments as those described in Sec. 5.1.1 by using different data sizes.

Figure 15 shows the obtained results. The graph does not include the sentiment score because the considered methods do not have reasonable values like the previous results (average sentiment score = 0.03). The results show that the proposed method is better than the other methods (see Fig. 15). We assume that it reflects the time influence from the start of the analysis; therefore, it provides a better result. However, the proposed method is not as good as the others when we use a small data size (under 7000) because the data are not sufficient for the analysis. However, the proposed method is effective if there is sufficient accumulated data. Note that the accumulated data do not only include size data, they also include period data. Moreover, as you can see, the proposed method is stabilized when the content size is greater than or equal to 8000; however, even this data size is almost a tenth of the whole data size. This result shows that the proposed method is useful even in the case of a small data size.

The purpose of this experiment is to show that the proposed method is reasonable even when used for a small amount of data. The obtained results confirm this hypothesis, demonstrating that the proposed method is reasonable even if there is a small amount of data. As the data size increases, it becomes stable.

5.2. *Discussion of useful characteristics for Twitter analysis*

Here, we will discuss the characteristics that are useful for a Twitter content analysis. We choose the characteristics used in this study and discuss their effectiveness. These

characteristics are frequency, number of retweets, number of followers, time, numerical statistics of the post, users popularity, and sentiment information.

- Frequency: Frequency is a simple parameter to analyze popularity or trends on Twitter. If a specific keyword comes up frequently in posts, it attracts public attention. Therefore, the proposed method is based on frequency, idf, and content influence. In the experiment, frequency is always reasonable for small to large amounts of data (see Sec. 5.1.2). Considering these findings, we conclude that frequency is one of the most important and effective characteristics for analyzing Twitter content.
- Number of retweets: We can say that the number of retweets is the number of shares of posts. In previous research, the retweet count was an important characteristic to measure content influence [10]. However, in this research, we found that it is useful for analyzing even data that included nonretweeted posts. Although the number of retweeted posts considered in this study was very small, it affected the measurement of the content influence (see Sec. 4.2). Therefore, we assume that the retweeted posts form a small amount of data but are useful in a general environment.
- Number of followers: The number of followers represents a user's popularity and the spreadability of contents [4, 10]. As with previous research, our experimental results show that the number of followers is useful for analyzing Twitter content (see Sec. 4.2). Because every Twitter post lists the author's follower, there is no sparseness of data. Therefore, it can be an effective characteristic to analyze Twitter content if used effectively.
- Time: Time is the most important factor in a Twitter search, because Twitter provides content research in a time-descending order. This implies that the currency of time is an important factor in Twitter [10]. Our study also shows that the time factor is helpful for analyzing Twitter content (see Sec. 4.2). It is useful to use accumulated time data, and the use of time with other characteristics is better than the use of only the time factor (see Sec. 4.2).
- Numerical statistics: In this study, we performed numerical statistical analyses such as tf–idf. These provided a reasonable result and were slightly better than the frequency measurement and PageRank methods, even when the Twitter contents had a monotonous keyword count. Because Twitter allows posts of 140 characters or less, the length restriction of Twitter ensures that there is no difference in keyword frequency and the document length between posts. Experimental data show that the average number of keywords in documents is 1.11, and the standard deviation is 0.37. We assume that term specificity (i.e. idf) is the reason why tf–idf is better than the frequency measurement and PageRank methods. The frequency measurement method considers only the number of posts containing the keywords, and the PageRank method is also based on the frequency of keywords. However, tf–idf considers not only the frequency of posts but also the term specificity.

Therefore, we conclude that numerical statistics is useful to analyze Twitter content with term specificity (i.e. idf).

- User's influence: We performed experiments measuring a user's influence by using a modified PageRank method (see Sec. 5.1.1). The result of the user's influence has a moderate coefficient correlation, but it is not as good as the other methods except for the sentiment analysis. However, there are several types of user influence considered in the previous research [4]. User influence measurements should be reanalyzed with several types of user influence. However, measuring a user's influence also requires a large amount of data, so it is not compatible for an analysis with a small amount of data.
- Sentimental information: Previously, in this paper, we assumed that positive and negative sentiment analysis results are useful for analyzing Twitter posts. However, our results show that sentiment information is not helpful for analyzing Twitter content because human sentiments are very difficult to express as a formula. As the experimental results show, a sentiment expression in Twitter has many variables. For example, if a specific subject is popular in the real world, it can have a negative sentiment ratio to cause a lot of negative sentiments, because the public users have interest in it. But another subject being less popular previously could have positive sentimental ratio because only few users maintained positive sentiments about it, and the public has no interest of it. However, this ratio was not useful in this study, but it can be a useful factor if it is analyzed in detail.

6. Conclusion and Future Work

In this study, we proposed an evaluation method based on the influence for an analysis of Twitter content. The goal of the Twitter content analysis in this study is to be efficient with a small amount of data, and to find a useful factor to analyze Twitter content. We used the crawled Korean Tweet data and the user relation data from 1 July 2012 to 31 July 2012, extracting nine subjects related to the music domain and six subjects related to the movie domain. We proposed the use of three characteristics: the number of followers of the content author, retweet count, and currency of time. We compared the results of the proposed method with numerical statistics, user's influence and sentiment score. Our experimental results showed that the proposed method using the influence with an accumulated period performs slightly better than the other methods; moreover, the proposed method performs reasonably well for a small amount of data.

We discussed factors that are useful for analyzing Twitter on the basis of this study's experiments. We extracted frequency, followers, retweet, and time to analyze Twitter content with only its own value. Numerical statistics and user influence are also useful, but these require a large amount of data. The sentiment information are not useful for analyzing Twitter content because the sentiment information had many variables. In the future, we want to expand the proposed method using content

influence to other SNSs, and to develop a better evaluation method for sentiment analysis. We also plan to improve our methods for a small amount of content.

Acknowledgments

This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF 2012M3C4A7033346) and the Basic Science Research Program through the NRF funded by the Ministry of Education (2016R1D1A1B03936375). Further, we would like to express our sincere gratitude to Daumsoft. Doo-Kwon Baik is the corresponding author. Young-Gab Kim is the co-corresponding author.

References

1. A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, Sentiment analysis of Twitter data, in *Proc. Workshop Languages in Social Media*, 2011, pp. 30–38.
2. S. Asur and B. A. Huberman, Predicting the future with social media, in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology*, Vol. 1, 2010, pp. 492–499.
3. J. Benhardus and J. Kalita, Streaming trend detection in twitter, *Int. J. Web Based Commun.* **9**(1) (2013) 122–139.
4. M. Cha, H. Haddadi, F. Benevenuto and P. K. Gummadi, Measuring user influence in Twitter: The million follower fallacy, in *Proc. ICWSM 2010*, 2010, pp. 10–17.
5. A. Esuli and F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in *Proc. LREC*, Vol. 6, 2006, pp. 417–422.
6. A. Go, R. Bhayani and L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford University, 2009, pp. 1–12.
7. G. Golovchinsky and M. Efron, Making sense of Twitter search, in *Proc. CHI2010 Workshop on Microblogging: What and How Can We Learn From It?*, 2011.
8. Google, Where trends data comes from, https://support.google.com/trends/answer/4355213, 2015.
9. H. Kwak, C. Lee, H. Park and S. Moon, What is Twitter, a social network or a news media?, in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.
10. E. Lee, J.-D. Kim and D.-K. Baik, Method for measuring Twitter content influence, in *Proc. Int. Conf. Software Engineering and Knowledge Engineering*, 2014, pp. 659–664.
11. C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Vol. 1 (Cambridge University Press, Cambridge, 2008).
12. Melon, chart (2015), https://www.melon.com/chart/index.htm.
13. Korean Film Council, KOFIC — Box office (2016), https://www.kobis.or.kr/.
14. A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in *Proc. LREC*, Vol. 10, 2010, pp. 1320–1326.
15. O. Phelan, K. McCarthy and B. Smyth, Using Twitter to recommend real-time topical news, in *Proc. Third ACM Conf. Recommender Systems*, 2009, pp. 385–388.
16. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Inf. Process. Manage.* **24**(5) (1988) 513–523.

17. M. Song and M. C. Kim, RT^{2}M: Real-time Twitter trend mining system, in *Proc. Int. Conf. Social Intelligence and Technology (SOCIETY)*, 2013, pp. 64–71.
18. B. Sun, and V. T. Y. Ng, Analyzing sentimental influence of posts on social networks, in *Proc. IEEE 18th Int. Conf. Computer Supported Cooperative Work in Design*, 2014, pp. 546–551.
19. J. Teevan, D. Ramage and M. R. Morris, TwitterSearch: A comparison of microblog search and web search, in *Proc. Fourth ACM Int. Conf. Web Search and Data Mining*, 2011, pp. 35–44.
20. A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, in *Proc.*, 2010, pp. 178–185.
21. J. Weng, E.-P. Lim, J. Jiang and Q. He, Twitterrank: Finding topic-sensitive influential twitterers, in *Proc. Third ACM Int. Conf. Web Search and Data Mining*, 2010, pp. 261–270.
22. K. Xu, S. S. Liao, Y. Song and L. Liu, Mining user opinions in social network webs, in *Proc. Fourth China Summer Workshop on Information Management*, 2010.
23. S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Comput. Netw. ISDN Syst.* **30**(1) (1998) 107–117.
24. Twitter4J, Introduction (2015), http://twitter4j.org/en/index.html.
25. JSON.org, Introducing JSON (2016), http://http://www.json.org/.