DEIM Forum 2011 A1-4

Twitter-based TV Audience Behavior Estimation for Better TV Ratings

Shoko WAKAMIYA † Ryong LEE ‡ Kazutoshi SUMIYA ‡

† Graduate School of Human Science and Environment, University of Hyogo

1-1-12 Shinzaike-honcho, Himeji, Hyogo, 670-0092 JAPAN

‡ School of Human Science and Environment, University of Hyogo

1-1-12 Shinzaike-honcho, Himeji, Hyogo, 670-0092 JAPAN

E-mail: † nd09a025@stshse.u-hyogo.ac.jp, ‡ {leeryong, sumiya}@shse.u-hyogo.ac.jp

Abstract Twitter is now used as a back channel to TV programs by many crowds who would like to express their opinions to watching programs or just simply put on records of TV viewing, etc. In terms of the popular trend of viewing TV with Twitter, we can utilize the microblgging site to conduct a large-scaled TV rating much easily comparing to the conventional TV rating methods accompanying relatively a quite small number of households. In this paper, we propose a system to measure TV ratings by analyzing TV lifestyles of the public through the microblogging service, Twitter. In order to realize Twitter-based TV ratings, we first need to identify potential audiences of each TV program. In addition, we have to find out reliable TV-viewing tweets for measuring TV ratings more precisely, since TV-relevant tweets include irrelevant ones to TV-viewing. However, in spite of these difficulties, crowd-based TV rating method has many benefits such as near real-time monitoring, almost zero-cost, feasibly survey of massive public's opinion, etc. In this paper, we present an estimation method of TV-viewing for tweets much definitely deciding if an audience is viewing TV by means of a learning-based message analysis. In the experiment, we describe our preliminary exploration result to distinguish tweets of TV-viewing from the other solely TV-relevant ones.

Keyword TV Rating, Microblog, Social Networking Service

1. Introduction

Due to the explosive growth of social networking sites led by Facebook[1] and Twitter[2], enormous number of people all over the world have actively been posting their real-time updates regarding what they are doing and what they are having in mind with their whereabouts. Interestingly, lots of crowds are utilizing these services as a back channel to TVs while watching TV shows. In fact, in NewTeeVee conference[3] held on Nov. 2010, an amazing fact concerning Twitter messages (called tweets) and TVs was presented: averagely 90 million tweets written per day are related to TV shows, and Twitter peak times are happening simultaneously during on-air TV prime time [11]. Like this, there is an increasingly popular trend that the combination of Twitter and TV programs goes well together for TV audiences who hope to share their opinions and sentiments for TV programs with others including TV broadcast stations. In addition, to measuring TV ratings is useful not only for audiences, but also for TV broadcast stations and even for advertising companies. Specifically, in the side of TV stations, they eagerly may want listen to opinions on their contents from a broader range of audiences. On the other hand, audiences would also like to often participate in the TV program by expressing their opinions or thoughts directly to the content providers. Accordingly, in terms of conventional TV viewing surveys, social media must be a valuable source to gather much bigger and wider audiences rating, with less additional costs to those selected participants who have worked for the conventional TV ratings.

In fact, the current TV ratings in the USA and Japan are measured on the basis of Nielsen ratings, which were developed by Nielsen Media Research [7] a long while ago. The developed method has measured TV ratings based on three different ways: first, "Set Meter" which is a method using an electronic device to monitor what TV programs the selected homes are viewing. Second, "People Meter" is the other method using a specially designed remote controller to recognize the members of a household who are watching the TV. Lastly, "Viewer Diary" which is audiences' self-recording on paper-based questionnaires about what they have watched individually. As for samples of the survey, because the current methods have already selected households, it is only necessary for analysts to observe those samples.

However, the advances in Internet and mobile technologies have been causing drastic changes in our media consumption patterns. For instance, we can increasingly view TV programs not only on our home television sets, but also on multiple platforms such as computers, mobile phones, or even automobile navigation systems. Furthermore, we also have numerous TV channels including the terrestrial analog or digital broadcasts, CATVs, satellite TVs as well as the expanding online video or TV sites. For instance, in the case of Japan, 230 different TV program's broadcast channels are available around the country. We can watch almost anything we want at anytime and from anywhere. The habit of watching TV will continue as an important part of our media life, despite the emergence of diverse media that are gradually stealing people's attention from the TV. As aforementioned, today's evolving media ecology is becoming more and more complicated. In fact, we can watch TV from places other than our homes. We can even watch TV programs on prerecorded shows, time-shifted replays, and on the Internet over the limitations of time and space. Amidst this complexity, conventional rating methods would not suffice.

In this paper, we propose a novel TV rating method considering the recent technical evolution, the diversification of crowds' media consumption styles, and crowd opinions. For this, we utilize the audiences' media-life logs over microblogging sites. In particular,

[1] http://www.facebook.com
[2] http://twitter.com/
[3] http://events.newteevee.com/live/10/

we focus on Twitter where many tweets concerning TV programs are being posted with additional tags of where and when audiences post their logs. However, this site was not designed for this specific goal to collect the TV-related Twitter messages identifying those that are relevant to TV programs, which are the target of this work. Therefore, we need to confirm the tweets to a particular topic in order to filter out other tweets from topics outside our interest. Especially, we present a semantic linking between tweets relevant to TV programs. In addition, because the topics about TV programs are spreading over Twitter users, non-audience might write TV-relevant tweets. In order to achieve our goal; to measure TV viewing rates by dealing with crowds' voices, we should distinguish irrelevant tweets to TV-viewing from relevant ones. However, it would be difficult to specify keywords to identify all the relevant tweets to TV-viewing manually. Therefore, we propose a method to estimate these TV-viewing tweets by utilizing a learning algorithm. By this method, we could conduct TV ratings reflecting crowd's voices in a real world.

The remainder of this paper is organized as follows: Section 2 describes current TV ratings and our motivation of this work, and reviews some related work. Section 3 describes a method for detecting TV audiences and estimating their behavior to figure out if they are truly watching TV programs. Section 4 illustrates experimental results using lots of tweets obtained from Twitter. Section 5 concludes this paper with further work.

2. TV Ratings using Massive Crowd Voices

2.1 Conventional TV Viewing Rates

Before we present our novel TV rating method, we review the most popularly adopted TV rating method. In general, TV ratings are known as a critical indicator to represent the value of TV program and its broadcast station. The current TV viewing rates have been measured on the basis of Nielsen's method [7] which were developed by Nielsen Media Research in 1950s and have applied in three ways as follows.

- **Set Meter**: It records what TV programs in selected homes are viewing by means of electronic devices connected to TV sets. The collected viewing logs are transmitted during nights or in real time to the Nielsen center or other media research companies.
- **People Meter**: It is able to recognize which members of a household is looking at TV programs by selecting one of buttons on a specially designed remote controller, eventually enabling analysts to survey various demographic groups such as younger vs. older generations.
- **Viewer Diary**: The oldest way is based on an audience's self-recording on a questionnaire paper about what s/he had watched on TV.

These three ways which have been conducted towards a quite limited the number of households due to the following reasons; only a quite small number (at most, a few hundreds of static households for a city) of randomly selected households become to participants in the daily survey for a period of time due to practical constraints of cost and taking time. Statistically, this has been a reasonable method, while many of audiences might be sometimes incepted by an intended TV programs ranking. Consequently, this well-established method dominating most TV ratings have become a de-facto standard.

Obviously, Twitter-based TV ratings have both advantages and disadvantages comparing to Nielsen ratings. We summarized the differences between the conventional statistical methods and our proposed method as shown in Table 1.

Table 1 Comparison of Nielsen ratings and our proposed TV ratings based on crowd-source

	Nielsen TV ratings	Twitter-based TV ratings
Scale of audience	Small	Huge
Participants	Randomly selected households	Anybody
Unit	Home-centric	Person-centric
Audience	Participants paid	Volunteers
Consciousness	Conscious	None or Unconscious
Period	Daily-based	Nearly real time
Place	Inside of house	Anywhere
Cost	Paying to the participants	Almost zero
Coverage	Local / National	From Local to Global
Methodology	Statistics on sample groups	Analysis of media logs transmitted by means of microblogs
Strength	•Well-established •Clear audiences' characteristics (age, sex, and so on)	•Massive quantity •Diversity •Opinions and sentiments •Location information
Weakness	•Small quantity •Less diversity •Hard to get opinions and sentiments	•Noises and fluctuations in descriptions •Unidentified audiences

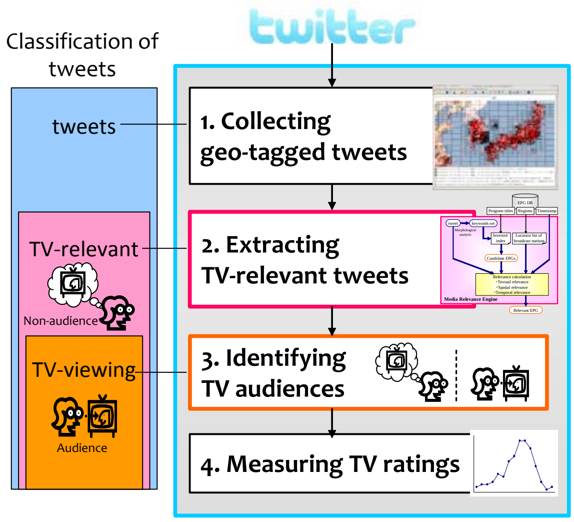

Figure 1 Concept of crowd-based TV rating system

2.2 Crowd-based TV Rating System

In this work, we aim to establish a crowd-based TV rating system to exploit the crowd's TV lifestyles for rating TV programs and looking into audiences' opinions. Especially, in this paper, we focus on the first and the most critical function to find out audiences from Twitter determining whether the found users are really watching TVs. Specifically, we have to detect potential TV audience as shown in Figure 1. For this, by analyzing tweets, we explore TV-viewing tweets. We consider that tweets by audiences who are watching TV programs would be written using words directly showing their current behavior, such as "watching," "watched," "viewing," and "viewed." Therefore, we regard the behavior of people using these words as "watching."

2.3 Related Work

In recent years, Twitter has been attracted increasing attention from researchers in various areas. Especially, lots of them are focusing on the characteristic of Twitter messages which can reflect phenomena in real world rapidly; almost real time. Mathioudakis et al. [5] have presented TwitterMonitor, a system identifies emerging topics like trend on Twitter in real time and synthesizes a description of each topic. Users can interact with this system by ordering the identified trends using different criteria and submitting a description concerning each trend.

Furthermore, the compatibility of Twitter with media has been aggregated. O'Connor et al. [8] compared the measures of public opinion from polls with ones from the analysis of tweets. Diakopoulos et al. [1] demonstrated an analytical methodology including visual representations and metrics that aid in making sense of the sentiment of social media messages around a televised political debate. In this paper, by finding tweets relative to TV watching, we estimate the public TV viewing rates. Sawai et al. [10] have proposed a method to recommend TV programs based on relations among users over social networking. Wakamiya et al. [12] have presented a method for rating TV programs and shared videos by grasping crowd's media lifestyle from their geo-tagged Twitter messages.

3 Twitter-based TV Ratings

3.1 Acquiring Tweets from Crowd

In this work, we target to measure TV ratings across in Japan. Specifically, in order to conduct a survey not bounded to home environments of the usual method adopted by the Nielsen rating, we are dealing with geo-tagged tweets which have information on the geographic locations of the writing users, specifically tagged with user ID, timestamp, location (lat. long.,) and textual message as shown in Table 2. By using these geo-tagged tweets, we can measure local TV rating which as one of our ultimate goal. In order to effectively collect such specific type of tweets, we utilize a tweet aggregation system developed in our previous work [4]. Since the system is not the main our focus of this paper, we do not describe the detail here.

Table 2 Example of geo-tagged tweets

User ID	Timestamp	Location		Textual Message
		Latitude	Longitude	
suren	Wed, 01 Sep 2010 00:00:39 +0000	41.080333	141.251414	I'll go to bed after this news has finished.
kjoyw	Wed, 01 Sep 2010 00:00:25 +0000	41.080333	141.251414	I check your tweet just now. At this time yesterday, I was drinking beer in an Okinawan restaurant.
yuuri	Wed, 01 Sep 2010 00:07:30 +0000	40.039021	140.994441	I completed half of my work.
bekaz	Wed, 01 Sep 2010 00:15:51 +0000	40.126901	140.306252	There is a person who well summarized Harvard University's Justice with Michael Sandel.

3.2 Detecting TV-relevant Tweets over Twitter

In a geo-tagged tweet database, various tweets which are written by enormous number of crowds are stored in a similar manner. In order to utilize these tweets as a source for TV ratings, we need to distinguish TV audiences from other normal Twitter users as illustrated in Figure 2. Of course, not every audience is always writing only about TV programs. They can also be non-audience and normal users. Thus, we dynamically detect TV relevant users by measuring the degree of relevance of their tweets to TV programs.

3.2.1 Extracting TV-relevant tweets using hashtags

For identifying relevance between tweets to the most relevant program, we approach two different levels of identification processes. First, a primitive and essential step's approach is to extract tweets which have already been made a connection with a TV program by means of hashtags like '#' and a string which utilized for aggregating same topic's tweets. Therefore, we look up a prepared hashtag list in the local hashtag database which includes hashtags and the source information such as a name of broadcasting station and a title of TV program as shown in Table 3. However, hashtag-based linking from a tweet to a relevant TV program will eventually suffer from lack of relevance enough to measure scores of TV programs because audiences are required their effort not only to manually write such hashtags into the writing tweets, but also to previously know the correspondence between a hashtag and TV.

Table 3 Example of hashtags used for TV stations or programs

Hashtag	TV station	TV program
#nhk	nhk	-
#tbs	tbs	-
#tvtokyo	tvtokyo	-
#etv	nhk_edu	-
#keion	-	Keion!
#precure	-	Precure
#gegege	nhk	Wife of gegege
#ryomaden	nhk	Ryomaden

3.2.2 Extracting TV-relevant tweets based on EPG

In general, for many people, it would be a common way to write just a title of TV program or a few keywords representing TV program. In addition, for dealing with lots of TV programs broadcasted in Japan, the hashtag-based method would not be sufficient. In other words, we cannot ignore such freely written texts which are connected to much more hidden TV audiences of huge variety of TV programs. Therefore, as the next step, we need to examine the relevance between tweets and possibly relevant programs from those raw and short texts. For this purpose, we use an Electronic Program Guide (EPG), which typically provides people with scheduling information for current and upcoming programs as shown in Table 3. Then, we compute the relevance between tweets and EPGs. For these, we have designed a Media Relevance Engine in our previous work [13]. In the engine, the relevance is calculated by means of textual, spatial, and temporal distances. We describe the detailed methods of each relevance as follows:

· **Textual Relevance**

In order to find a corresponding EPG item relative to a tweet about a TV program, we applied a words-based similarity computation: both sides are textual message. In the estimation of the correspondence, we compute it with the following formula based on the Jaccard similarity coefficient [2],

$$textual\_relevance = \frac{|mp(tw_i) \cap mp(e_j.title)|}{|mp(tw_i)| + |mp(e_j.title)| - |mp(tw_i) \cap mp(e_j.title)|} \times \sum_{k}^{|mp(e_j.title)|} \frac{1}{df(k)} \quad (1)$$

where tw_i is a tweet, e_j is an EPG item, $e_j.title$ is the title in the EPG item, and mp is a morphological analysis function where the output consists of nouns found in the given message. df is a function that calculates document frequency. Each tweet should be compared

Table 3 Example of EPGs

Region	Station	Date	Time Start	Time End	Title	Genre
CATV Tokyo area J:COM Tokyo (Suginami)	NHK General Tokyo	Sep. 1, 2010	0:00	0:15	News and weather information	news
CATV Tokyo area J:COM Tokyo (Suginami)	NHK General Tokyo	Sep. 1, 2010	1:05	1:50	Chase! A to Z	documentary
CATV Tokyo area J:COM Tokyo (Suginami)	NHK General Tokyo	Sep. 1, 2010	1:50	2:00	Scoop! Contributed video clips (Tokudane! Toukou DO-ga)	talk show
CATV Tokyo area J:COM Tokyo (Suginami)	NHK General Tokyo	Sep. 1, 2010	2:00	2:45	Try and convince (Tameshite Gatten)	talk show / lifestyle

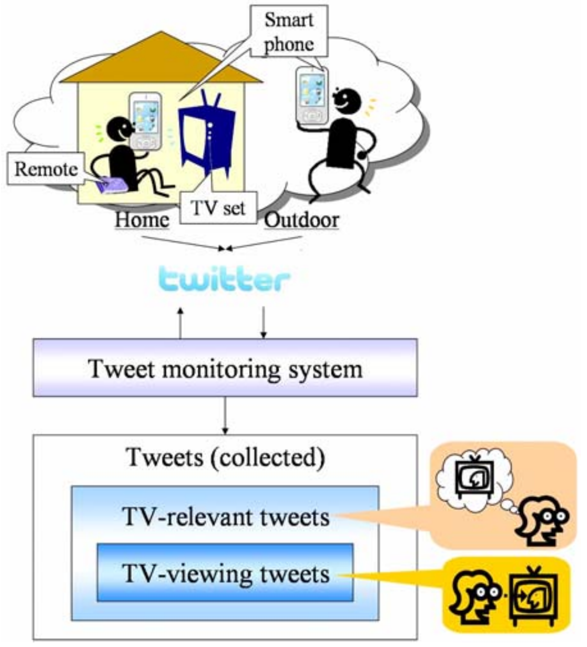

Figure 2 Categorization of collected tweets

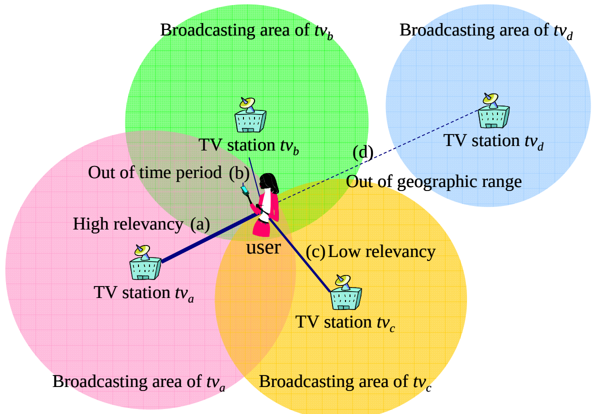

Figure 3 Filtering TV programs based on Spatial and Temporal Distance

with all the program titles in the local EPG database. For the rapid searching for seemingly relevant EPGs we use an inverted index [15] to reduce the number of calculations required for determining relevance between a tweet and program titles in comparison to directly using the EPG database wherein the total of possible combinations would be enormous. Then, in order to detect relevant EPGs related to titles of TV programs, we applied the formula (1) in the computation. In the formula, with the *df*, we also considered the frequency of keywords of EPGs' titles. For example, keywords that are frequently used in EPGs such as "news," "drama," and "sports" should have less weight since these generic terms would retrieve many unrelated EPGs.

・ Spatial and Temporal Distance

According to EPG items in the local EPG database, the same titles of EPGs are often found, because some TV programs can be broadcasted repetitively by multiple stations. In this case, we should identify the station that broadcasted the program at the time of tweet occurrence. The number of TV programs extracted by the inverted index usually corresponds with many different local stations. However, a user can exist at a place in a given moment so that a TV-relevant tweet should be matched to one of the possible local stations. Therefore, we should consider the physical distance between the location where a tweet is posted and that of the station that broadcasted the TV program. Because specific locations of stations are not included in EPG items, we roughly estimate their locations based on 'region' attributes of the EPG items using Google maps API[4]. For this, we use the stations' location list that was generated beforehand. Then, we calculate distances between a location where a tweet was posted and each station, and the station that has the minimum distance is selected.

$$\begin{aligned} &spatial\_distance \\ &= \log(euclid\_dist(tw_i.location, e_j.tv\_station_k.location)+1) \end{aligned} \tag{2}$$

There is also an important consideration regarding the tweet posting time. Usually, we can think that TV-relevant tweets may be written near the actual on-air time. For instance, audiences may write a lot of tweets during or just after a popular drama. Sometimes, before a very popular sports program such as the World Cup, many tweets may occur far before the actual on-air time. Therefore, as regards the relevance between tweets and TV programs, the time elapsing between them is also an important factor.

For instance, as drawn in Figure 3, a user is a location in the middle of a city and there are four different broadcast stations around there. But only three stations tv_a, tv_b, and tv_c are accessible from the location of the user. If a tweet written by this user is matched with some program information broadcasted from the surrounding four stations, we can think that the user's message can be to these programs. However, the station tv_d cannot support this assumption, since it is out of the period. Furthermore, in terms of broadcasting time, it is likely that the programs broadcasted in the nearly same time range with the written time would be desirable. Therefore, we computed the relevance of tweets to find out relevant on-air programs in the respects of textual relevance, spatial and temporal distances.

$$temporal\_distance(tw_i, e_j) = \begin{cases} 0 \\ \quad \text{when } e_j.start\_time \le tw_i.timestamp \le e_j.end\_time \\ \log(e_j.start\_time - tw_i.timestamp + 0.1) \\ \quad \text{when } tw_i.timestamp < e_j.start\_time \\ \log(tw_i.timestamp - e_j.end\_time + 0.1) \\ \quad \text{when } e_j.end\_time < tw_i.timestamp \end{cases} \tag{3}$$

・Final Relevance

Based on the above criteria, we computed the final relevance for finding out TV-relevant tweets using the following formula:

[4] http://code.google.com/intl/ja/apis/maps/.

$$relevance\_score(tw_i, e_j) = \frac{textual\_relevance(tw_i, e_j)}{(spatial\_distance(tw_i, e_j)+1) \times (temporal\_distance(tw_i, e_j)+1)} \quad (4)$$

3.3 Estimating Reliable Audiences

TV-relevant tweets extracted in Section 3.2 would have written by both audience and non-audience. In order to achieve better TV ratings, we should extract tweets written by TV audiences from these mixed tweets. In other words, we need to identify reliable audiences who are really watching TV programs. We assume that these audiences would utilize specific words concerning TV-watching. Therefore, we need to prepare the words which are often used by audiences during watching TV programs. For this, we can obtain the words by two ways; (1) specifying manually such as "視聴," "見る" and (2) utilizing a supervised machine learning algorithm. However, in the case of (1), it is not easy to expect such keywords which can clarify the status of TV watching. So, in this work, we got words by the supervised machine learning algorithm automatically.

However, it is not easy to expect such keywords which can clarify the status of TV watching. Therefore, we decided to use a supervised machine learning algorithm and classify TV-watching tweets and the other. As for the classification algorithm, we apply Naïve Bayes Classifier [9] because it is simple and often used as a baseline method but we can expect to provide better performance. Furthermore, Naïve Bayes Classifier is practically used for filtering spam mails; the algorithm might be helpful for our targeting goal. We show the basic formula of Naïve Bayes Classifier as follows:

$$P(cat \mid tw) = P(w_1 \wedge ... \wedge w_k \mid cat) = \prod_i P(w_i \mid cat) \quad (5)$$

where *P* is a function which calculates the occurrence possibility, *tw* means a tweet. *cat* is a category; TV-viewing or other, and w_i means a word. *tw* can be also expressed as bag-of-words. In the experiment described in section 4, we utilized bag-of-words segmented by a Japanese morphological analyzer MeCab [6]. Here, a tweet is not needed to parse because we attempt to extract words related to TV-viewing regardless of their word class. In the formula (5), *P(cat|tw)* returns a possibility which a tweet *tw* is generated when a category *cat* is given.

$$P(w_i \mid cat) = \frac{F(cat, w_i)}{\sum_{w_j \in Voc} F(cat, w_j)} \quad (6)$$

In the formula (6), *F* is a function which returns the frequency of occurrence, hence $F(cat, w_i)$ returns the total number of a word w_i in a category *cat*. *V* is a set of all vocabulary occur in training data.

$$cat_{map} = \arg\max_{cat} P(cat \mid tw) = \arg\max_{cat} P(\log P(cat) + \sum_i \log P(w_i \mid cat)) \quad (7)$$

$$P(w_i \mid cat) = \frac{F(cat, w_i) + 1}{\sum_{w_j \in Voc} (F(cat, w_j) + 1)} = \frac{F(cat, w_i) + 1}{\sum_{w_j \in Voc} (F(cat, w_j) + |V|)} \quad (8)$$

After matching between TV-related tweets and the TV-viewing words, we estimate the rest tweets are written by audiences only "noticing" the TV program.

4 Experiment

4.1 Experimental Dataset

In order to achieve our purpose to rank TV programs by means of Twitter users, we prepared a dataset for a period between September 1–30, 2010: tweets that occurred in that period in Japan, and EPGs of all TV stations (except CS satellite broadcast) in Japan. In that period, we collected 6,276,769 geo-tagged tweets, which were all mapped onto location points on a map. However, it was still burdensome to use this tweet dataset in our preliminary test. For the practical findings of TV-relevant tweets, we empirically chose tweets whose relevance to TV watching was seemingly higher using the prepared hashtag lists (for on-air TVs and on-line videos) and a set of filtering terms such as "テレビ," "TV," "てれび"–Japanese expressions for "television," and "視聴," "番組," "見てる," "見ている"–expressions for "watching" or "viewing." By filtering using these terms, we could obtain a reduced tweet dataset (119,575 tweets, about 1.9% of the collected dataset). These potential tweets were written by 33.392 distinct users (on average, 3.58 tweets were made per a user.)

By using these dataset, we conducted an experiment concerning identification of TV-relevant tweets, and estimation of TV-viewing tweets.

4.2 Experimental Result

In our proposed system, we first find out TV-relevant tweets and then distinguish TV-viewing tweets from those of TV-relevant. The experiment for detection of TV-relevant tweets had been conducted in our previous work [12] as shown in Table 5. This is the ranking result based on popularity score of each TV program. In table 5, '*type*' means a type of extraction method; whether a program was extracted using hashtag or epg. '*#tweets*' is the number of TV-relevant tweets of each TV program, '*#users*' is the number of users who posted the TV-relevant tweets, and '*pop.*' shows a logarithmic value of popularity score computed by the following formula:

$$popularity(e_j) = \log(\sqrt{\#tweets \times \#users})$$

Next, in order to estimate TV-viewing tweets, we conducted the experiment using 300 TV-relevant tweets written in July 2010 as a training data; 150 of 300 TV-relevant tweets are included in TV-viewing ones. The number of vocabularies in the training data was 1,551. Here, we split a textual message into words using Japanese morphological analyzer, Mecab [6] by inserting space between each word. For confirming an accuracy of the TV-viewing crowd detection engine, we prepared 100 tweets written in September, 2010 as a test data. We distinguished "TV-viewing" tweets from "TV-relevant" tweets manually, 50 tweets respectively, in advance. In the case of a test tweet "I'm watching closeup Twitter," the possibility score which the test tweet *test_tw* will be estimated as a tweet by an audience, *log P(audience|test_tw)* was

Table 5 TV program ranking based on TV-relevant tweets

rank	Titles of programs	type	#tweets	#users	pop.
1	Keion! (anime)	hashtag	978	141	3.64
2	Wife of gegege (drama)	hashtag	138	59	2.84
3	Ryoma's history (drama)	hashtag	76	47	2.61
4	Precure (anime)	hashtag	145	27	2.51
5	If I become a prime minister (talk show)	epg	45	44	2.47
5	Keion! "Examination" (anime)	epg	45	43	2.46
6	Leading actor ―setting beaty salon― (comedy)	epg	32	30	2.23
7	Revolution TV (talk show)	hashtag	59	22	2.23
8	BGM for working "Yamaguchi, Kyushu where Ryoma loved"	epg	31	26	2.16
9	Sengoku BASARA 2 #12 (Anime)	epg	26	25	2.11
10	Miyane'd house (talk show)	epg	26	23	2.07
11	IT white box II	epg	25	23	2.06
12	world_business_satellite	hashtag	57	15	2.05
13	Honmadekka TV	epg	26	22	2.05
14	Keion! "After school" (anime)	epg	26	21	2.03
15	Majisuka school	epg	22	22	2.01

Table 6 Examples of TV-viewing tweets

Textual messages	Estimated category
クローズアップついったー**見てるなう**	TV-viewing
ドラえも～ん！隣で弟が**見てる**。	TV-viewing
クルマのなかで**テレビ見てるなう**。	TV-viewing
あれ、また**テレビ**に要芽姉が**出てる**よ。	TV-viewing
こんどは、「けいおん」を**見てる**? アイスクリーム食べながら?	TV-viewing
NHK教育テレビ『10min.ボックス』録画のを**見た**。テーマは「CMを作る」で、インターネット広告。遅い時間帯の放送だけど、過去のも中身は興味深いので、今後も見ていく。10分という短さもGoooD。	TV-viewing
オカルト学院**見てる**けど、なにが面白いのかよくわからん (*´Ａ`*)	TV-viewing
おはよー！**朝**めざましテレビ**見て**、superflyかっこよかった～～	TV-viewing
受験みたいだよ　**今**、**見てる**けど	TV-viewing
けいおん！！21話**見てるなう**	TV-viewing

Table 7 Examples of TV-relevant tweets

Textual messages	Estimated category
@sinkzink 周りからクレーム来て、別垢にしたんですよー。しかもあんまりテレビを**見てない**という。。。。電車の中で**見たい**番組が**ない**っす。	TV-relevant
@maharad66マジですか～**見逃し**ました。youtubeで**見れる**かな～	TV-relevant
@sgtmm0lf 何時のテレビ？？**録画**する(ー`дー´)ｷﾘｯ	TV-relevant
@toko_al おつかれさまでした～！でも、ちょうどバタバタしていて番組**見逃し**ちゃいました…	TV-relevant
ありがとうございます。**見まっす**！ RT @michael072 **今夜**8時かららしいです。　RT @Texas4619CRANK: そういえば、所さんの番組にルアマガが出るの**今日**でしたっけ？	TV-relevant
誰も**見てない**YouTubeビデオ大集合　http://bit.ly/9Ssm0	TV-relevant
※けいおん！、けいおん!!**未視聴**	TV-relevant
24時間テレビで辻仁成がマラソンやったら絶対に**観る**。	TV-relevant

-67.02, and the one it will be estimated as a tweet by a non-audience, *log P(non-audience|test_tw)* was -70.67. Specifically, the possibility which "watch" is used by audiences; *P("watch"|audience)* was 0.00033, and the one which indicating progressive "てる" in Japanese is used by audiences likewise; *P(progressive form ("-ing")| audience)* was 0.050116. This possibility was calculated by the formula (5). On the other hand, the possibility the word "watch" is used by non-audiences; *P("watch" |not-watch)* was 0.00030, and the one which indicating progressive "てる" in Japanese is used by non-audiences likewise; *P(progressive form ("-ing")| non-audience)* was 0.00274. This possibility was calculated by the formula (8). Both words are higher possibility classified into "watch." Hence, this tweet was correctly classified into "watch." In Table 6, we show examples of tweets correctly estimated into the category "TV-viewing."

We also evaluated the accuracy by calculating the ratio of the number of tweets classified correctly in the total number of test data. In this evaluation, the accuracy of the engine using naïve bayes classifier was 70%. To be specific, 39 of 50 TV-viewing tweets (78%) were identified as TV-viewing correctly as shown in Table 6. On the other hand, 31 of 50 TV-relevant tweets (62%) were identified as TV-relevant correctly as shown in Table 7. Consequently, although the classification accuracy of tweets which consisting of a few words or many words became low. Therefore, these tweets needed to be taken into additional consideration. We were successfully able to distinguish TV viewing tweets from TV-relevant tweets as illustrated in Tables 6 and 7.

5 Conclusion

In this work, we have attempted to measure better TV ratings based on tweets as crowds' voices. In this paper, we first gathered geo-tagged tweets using the geographic tweet monitoring system and extracted tweets related to TV programs (TV-relevant tweets) by finding out potential TV audiences by means of hashtags and EPGs from collected tweets. In this paper, we focused on the difference between audiences and non-audiences for realizing better TV ratings. For this, we presented an estimation method of TV-viewing for tweets much definitely deciding if an audience is viewing TV by means of a learning-based message analysis. In the experiment, we utilized Naïve Bayes Classifier as a baseline method and showed our preliminary exploration result to distinguish tweets of TV viewing from the other just TV-relevant ones successfully. As the next step of this work, for realizing Twitter-based TV ratings, we are going to explore the opinions or sentiments of TV audiences based on TV-viewing tweets.

In our future work, we will seek much deeper crowds' TV viewing lifestyles by analyzing opinions and sentiments for tweets. On the basis of TV-viewing tweets considered sentiments of messages such as negative or positive, we should rank TV programs in terms of different scales for short or long periods or for local or global regions and show results. In this point, although our method is difficult to determine TV programs' distribution types such as air or DVDs precisely, we might consider it by giving a weight for the temporal distance in part. Furthermore, we attempt to measure cross media ratings by extending our method for various media such as radios, news papers, magazines and movies as well as TV programs.

Acknowledgment

This research was supported in part by a Grant-in-Aid for Scientific Research (B)(2) 20300039 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

[1] Diakopoulos, N. A., Shamma, D. A., Characterizing debate performance via aggregated twitter sentiment. In Proc. of the 28th international conference on Human factors in computing systems. pp. 1195-1198, 2010.

[2] French, J. C., Powell, A. L., and Schulman, E.: Applications of approximate word matching in information retrieval: In Proc. of the sixth international conference on Information and knowledge management (CIKM '97), 1997.

[3] Jansen, B. J., and Zhang, M., Sobel, K., and Chowdury, A.: "Twitter power: Tweets as electronic word of mouth." Journal of the American Society for Information Science and Technology 60(11), pp. 2169-2188, November, 2009.

[4] Lee, R., Wakamiya. S, and Sumiya, K.: Discovery of Unusual Regional Social Activities using Geo-taggged Microblogs," the World Wide Web Journal, Special Issue on Mobile Services on the Web, March 2011.

[5] Mathioudakis, M. and Koudas, N.: "TwitterMonitor: Trend Detection over the Twitter Stream," In Proc. of the 2010 international conference on Management of data (SIGMOD'10), pp. 1155--158, 2010.

[6] MeCab: http://mecab.sourceforge.net/.

[7] Nielsen Media Research: http://en-us.nielsen.com/content/nielsen/en_us.html

[8] O'Connor, B., Balasubramanyan, R.: Routedge, B., and Smith, N. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proc.of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010).

[9] Rish, I.: An empirical study of the naive Bayes classifier, In IJCAI-01 workshop on Empirical Methods in AI, 2001.

[10] Sawai, R., Ariyasu, K., Fujisawa, H., Kanatsugu, Y.: TV

[11] Tsotsis, A.: “There Are 90 Million Tweets Per Day, And A Lot Are TV Related,” http://techcrunch.com/2010/11/10/twitter-tv/, Nov 10, 2010.

[12] Wakamiya, S., Lee, R., and Sumiya, K.: “Crowd-powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs,” In Proc. of 2nd International Workshop on Social Networks and Social Media Mining on the Web (SNSMW), April, 2011 (to appear)

[13] Wakamiya, S., Lee, R., and Sumiya, K.: “Towards Better TV Viewing Rates: Exploiting Crowd's Media Life Logs over Twitter for TV Rating,” In Proc. of the 5th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2011), Feb. 2011.

[14] Zhao, D. and Rosson, M. B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In Proc. of the ACM 2009 International Conference on Supporting group work. pp. 243-252, 2009.

[15] Zobel, J. and Moffat, A.: Inverted files for text search engines. ACM Computing Surveys (CSUR 2006), Vol. 38, Issue 2, Article 6 (2006).