2016 IEEE 16th International Conference on Data Mining Workshops

SmartVideoRanking: Video Search by Mining Emotions from Time-Synchronized Comments

Kosetsu Tsukuda, Masahiro Hamasaki, and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
Email: {k.tsukuda, masahiro.hamasaki, m.goto}@aist.go.jp

***Abstract*—Many people search for and watch videos on video sharing Web sites, where users input a query and rank videos on the basis of metrics such as view count and rating. However, it is not always easy to find the desired video with a conventional search. One approach that enables users to more intuitively search for videos they desire is to index them according to emotions. Previous studies have used several predefined emotion categories, such as "fear" and "funny", for this purpose. However, viewers' emotions tend to be more diverse and specific. In this paper, we dynamically detect emotions in accordance with an input query and implement SmartVideoRanking, which enables users to search for videos on the basis of the detected emotions. We estimate viewer emotions from time-synchronized comments on videos and estimate the usefulness of each emotion by using support vector machine regression. Experimental results show that: (1) Spearman's rank correlation between the estimated usefulness scores and gold standard data was 0.7547; (2) emotions associated with videos vary from one query to another and it is therefore meaningful to detect emotions according to an input query; and (3) rankings based on viewer emotions enable users to browse videos that do not appear at the top of conventional search results. We also conduct a user study and demonstrate SmartVideoRanking's capability to search for videos.**

I. INTRODUCTION

Thanks to the popularization of video sharing Web sites, video recording devices, and video editing software, not only professional creators but also consumers can now easily create videos, known as user generated content (UGC), and make them accessible on the Web. This development has enabled us to watch videos on video sharing Web sites. Although the amount of uploaded videos is constantly increasing, the ways of searching for them remain limited. Users typically input a text query that corresponds to the video's title and/or tags, and the retrieved videos are then sorted or ranked according to metrics such as number of views and rating. However, it is not always easy for users to find videos they desire because they may have trouble correctly representing the intent of their search as a query [1], and UGC typically contains fewer textual tags and content descriptions than professional videos do [2].

One way to enable users to more intuitively search for videos they desire is to search on the basis of concepts and events (*e.g.*, "dog", "beach", "baseball", and "wedding reception"). There is a significant amount of literature on detecting concepts and events by using audio, visual, and textual features in videos [3]–[12]. Another way, which has recently become a popular research topic, is a search based on *emotions*. Such a method allows users to search for a video that belongs to an emotional category such as "joy" or "anger". In the past, the main target of emotion detection was movies [13], [14], while more recent studies have focused on emotion detection in UGC [15]–[17]. Different kinds of features have been used for video emotion detection, including visual and audio features [15], [16] and textual features [17], [18]. Note that the emotion categories of these studies were predefined (*e.g.*, six categories [17] and eight categories [15], [16]). However, viewers may have more diverse and specific emotions while watching a video. For example, when watching a video in which a person covers a song, they may feel that the singer's voice is "gentle" or they may feel "comforted". If videos were indexed on the basis of more diverse and specific emotions, users would be able to search videos more intuitively and the situations that could be distinguished could be greatly expanded, including ones such as "I am tired from work, so I want to search for a cover video by which *I will feel comforted*."

In light of the above, in this paper, we propose a video search system called SmartVideoRanking for detecting emotions from an input query rather than using predefined emotions. SmartVideoRanking shows the detected emotions to the user so that s/he can search various videos by clicking her/his emotion of interest. It generates a ranked video list according to the clicked emotion. To obtain viewers' emotional reactions to videos, we use comments on NicoNico[1], which is one of the most popular video sharing Web sites in Japan (described in Section II-A). On NicoNico, viewers can post comments at arbitrary temporal playback positions in the video while they are watching it. Because viewers' comments are the reactions to videos, in this paper, we assume any comment is related to an emotion and believe that such time-synchronized comments are effective resources for obtaining useful responses for searching videos on the basis of emotions.

Although video sharing Web sites feature a variety of uploaded videos, music is one of the most popular categories. Music videos include those in which a user has written the song and those in which a user dances to or covers other users' original songs [19], [20]. The latter sort are called *derivative works* [20]. Derivative works tend to have similar titles and tags: for example, videos that cover a song titled "Melt" are often titled something like "Melt Cover." This causes difficulty when searching using titles and tags. Hence, in this study, we focus on original music videos and their derivative works.

[1] http://www.nicovideo.jp

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.43

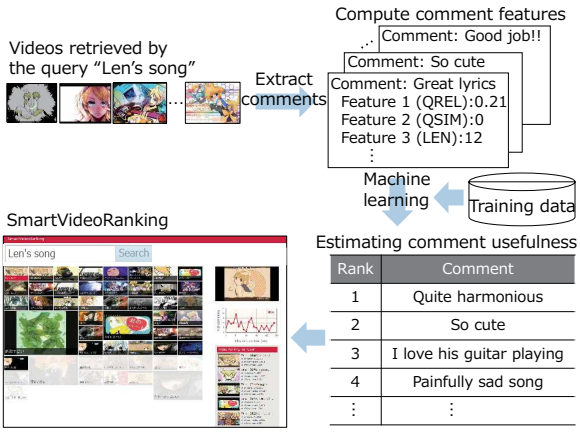

Fig. 1. Flow chart of our proposal.

Fig. 2. Interface of video player on NicoNico.

Figure 1 gives an overview of our proposal together with some examples. The query is "Len's song"[2]. The first step is to obtain videos related to the query. The second step is to extract comments, such as "Great lyrics" and "So cute", that were posted on the videos. The third step computes feature values for each comment. Here, each comment corresponds to an emotion. Features are categorized into 14 groups such as the query relevance (QREL) and query similarity (QSIM). The fourth step uses machine learning to estimate the degree of usefulness of each comment. In this example, the comment "Quite harmonious" is estimated to be the most useful. SmartVideoRanking is implemented on the basis of the estimated results.

The following are the research contributions of this paper:

- We introduce the concept of video searches based on query-dependent emotions.
- We propose features for automatically identifying useful viewer emotions for a given query by using time-synchronized comments posted on videos.
- We conducted experiments and found that: (1) Spearman's rank correlation between the estimated usefulness scores and gold standard data was 0.7547; (2) emotions varied from one query to another and it was therefore meaningful to detect emotions according to an input query; and (3) ranking based on viewer responses enabled users to browse videos that did not appear at the top of conventional search results.
- We implemented a video ranking system called SmartVideoRanking. Results of a user study with SmartVideoRanking showed that the participants found the system to be suitable for searching videos and easy to use, with most saying that they would like to use the system frequently.

The remainder of this paper is organized as follows. Section II gives an overview of our target video sharing service, NicoNico, and present related work. Section III presents an overview of SmartVideoRanking. Section IV describes the features used to estimate the usefulness of the emotions posted by viewers. Sections V and VI describe our experiments and user study, respectively. Finally, we conclude in Section VII with a brief summary and mention of future work.

II. BACKGROUND AND RELATED WORK

Let us begin by introducing the video sharing Web site NicoNico and its unique comment functions. After that we will describe related work pertaining to event detection and emotion detection in videos.

A. NicoNico

NicoNico is one of the most popular video sharing services in Japan. As of the end of July 2016, approximately 13.5 million videos had been uploaded. One of the most unique functions of NicoNico is its commenting system. On other video sharing services, such as YouTube[3], comments posted by viewers are displayed below a video, while on NicoNico, viewers can post comments at arbitrary playback positions in the video itself (hereafter, we call such comments *time-synchronized*), and the posted comments of all viewers are overlaid directly onto the video and synchronized to their playback positions. We illustrate how this is done in Fig. 2, which shows the interface of NicoNico. While a user is watching a video, s/he can input a comment such as "Great lyrics!" or "I love this song" in a text box below the video and post it on the video. The posted comments are overlaid onto the video and flow from right to left as the video plays. Hence, if many viewers think a scene is funny and post comments, many comments such as "lol", "haha", and "funny" are overlaid. As Yamamoto and Nakamura [17] and Yoshii and Goto [21] have pointed out, this function allows users to share their viewing experiences and feel as if they are watching a video concurrently with others, even though they are actually watching the video at different times. Moreover, when viewers post time-synchronized comments, they tend to more directly input what

[2]*Len* is the name of the *VOCALOID* described in Section II-A.

[3]http://www.youtube.com

they felt using shorter comments compared with comments on services such as YouTube because they post comments while they are watching a video. Thus, whole comments in and of themselves tend to be useful indicators of emotions, without having to extract phrases from comments. For this study, we focused on these properties of time-synchronized comments and tried to predict viewers' emotions from such comments.

One of the most popular categories on NicoNico is music. Users of NicoNico like to create their own songs by using singing synthesizer software called *VOCALOID* [22]. More than 140,000 original song videos created using VOCALOID had been uploaded to NicoNico as of the end of July 2016. Another popular activity is to create derivative works [20] such as covers and dance arrangements of original song videos. As of the end of July 2016, more than 590,000 derivative works had been uploaded to NicoNico.

B. Event Detection in Videos

There is a significant amount of literature about event and concept detection in videos as a way to index and retrieve videos (*e.g.*, to facilitate searching for videos that contain "birthday celebration" scenes on a video sharing service or "goal" scenes from soccer game video archives). The existing approaches can be classified into content-based ones and external-knowledge-based ones. Most of the content-based approaches use audio, visual, and textual features. For example, audio features were used to detect 25 concepts such as "beach" and "baby" [7] as well as to detect baseball events [9]. In terms of visual features, Liu *et al.* [8] detected human actions from user-generated videos on YouTube and Ekin and Mehrotra [4] did the same for soccer events. Textual features have been used to index lecture videos [3] and sports videos [12].

The external-knowledge-based approaches are more related to our study. One of the most popular resources of external knowledge consists of the scripts of TV programs and movies. Scripts can be used to detect the actions [6], [23] and appearances of specific characters in scenes [5], [23]. In recent years, external knowledge resources have become diverse as more and more social media Web sites come online. Tapaswi *et al.* [11] used plot synopses of TV programs and movies from Wikipedia[4] for their story-based video search. Event detection in TV programs by using Twitter[5] is also a popular research topic. Although NicoNico is a Japan-specific service, on Twitter, it has become popular to post comments (a.k.a. tweets) about TV programs while watching them. We can say that such comments are also time-synchronized. By using such comments, Shamma *et al.* [10] detected the most discussed events in videos of the first 2008 USA Presidential Debate and the Presidential Inauguration of Barack Obama. Van *et al.* [24] used them to detect soccer events.

All of the studies described in this subsection aim to index videos on the basis of events and concepts, whereas we aim to index videos on the basis of viewers' emotions. To achieve our goal, we use time-synchronized comments posted on videos by viewers and various features to detect useful emotions for searching videos.

[4]http://www.wikipedia.org
[5]http://twitter.com

C. Emotion Detection in Videos

Another way to semantically index videos is to detect emotions in videos. Similar to event detection, this approach can be classified into content-based ones and external-knowledge-based ones. Poria *et al.* [25] studied content-based approaches, wherein emotions were estimated using text, visual, and audio features. Recently, Jiang *et al.* [15] and Pang and Ngo [16] focused on user-generated videos on YouTube and detected video emotions such as "anger" and "sadness" on the basis of visual and audio features.

Regarding the external-knowledge-based approaches, the studies most closely related to ours are those that focus on using time-synchronized comments. Diakopoulos and Shamma [18] used Twitter to analyze the sentiment ("positive" or "negative") for each person appearing in a TV program. Yamamoto and Nakamura [17] proposed a method for classifying music video clips uploaded to NicoNico into six music mood categories such as "cheerful," "wistful," and "aggressive" by using a support vector machine (SVM) in which features were extracted from time-synchronized comments. More specifically, the features included adjectives and lengthened words in the comments as well as comments in chorus sections. Their method outperformed existing methods that use the lyrics and audio signals of songs. Nakamura and Tanaka [26] proposed a method for indexing and ranking videos according to four types of impression: positive, negative, happy, and sad. Their method required a manually created dictionary containing a few hundred regular expressions to match each impression and extracted impression information from time-synchronized comments.

In these studies, the detected emotions were predefined (*e.g.*, six [17], [25] and eight [15], [16] categories). In contrast, we do not predefine emotion categories; rather, we dynamically discover emotions from the input query.

III. System Overview

This section overviews the interface and functions of SmartVideoRanking. We encourage readers to watch the demonstration video (http://youtu.be/ZJgoB0ILdV4) to get a better understanding of our system.

A. Grid-style Emotion Interface

The user first inputs a text query and then SmartVideoRanking displays the top 50 emotions in terms of their usefulness for searching videos. Usefulness is estimated with the machine learning, and features are described in the next section. Each emotion is displayed with its representative video's thumbnail so that the user can get an idea of the mood of the retrieved videos. Hereafter, we call the set of an emotion and its thumbnail a *grid*. Figure 3 shows the interface for the query "Cover", with which the user wants to search for videos in which a person covers an original VOCALOID song. Fifty

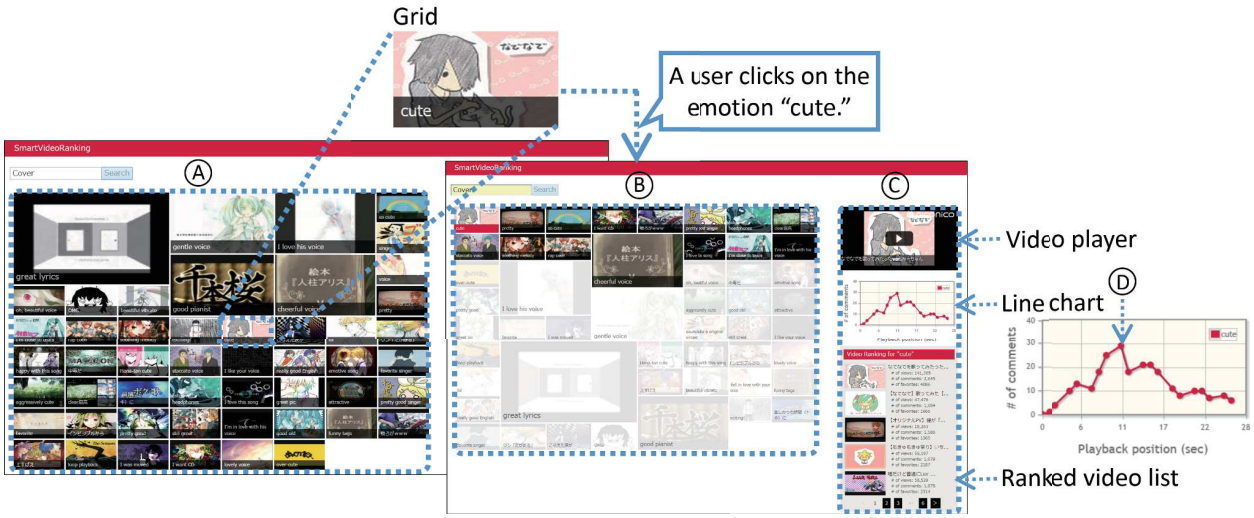

Fig. 3. Operation example of SmartVideoRanking.

emotions, such as "cute", "gentle voice", and "I am close to tears", are displayed together with the thumbnails (Ⓐ in the figure). Emotions are sorted in descending order of usefulness score: the emotion in the upper left corner has the highest score and the one in the lower right corner has the lowest score. In addition, SmartVideoRanking changes the thumbnail size and emotion's font size in accordance with the estimated usefulness score. To be more specific, the thumbnail size and font size are set to 404 x 236 pixels and 18 px, respectively, for the usefulness score ≥ 1.6; 266 x 154 pixels and 14 px for < 1.6 and ≥ 1.4; and 128 x 72 pixels and 12 px for < 1.4. This is to encourage the user to interact with the system even if s/he cannot decide on which emotion to select: the user can simply select the upper left grid.

When a user finds an interesting emotion, s/he clicks the grid. The right screenshot in Figure 3 shows the interface when the user clicks the emotion "cute". This subsection describes area Ⓑ in the figure. When a user clicks a grid, the grid moves to the upper left corner. Other grids are sorted in descending order of relatedness to the selected emotion. The relatedness between two emotions is computed on the basis of co-occurrence frequency in videos. Details are given in Section IV-E. This enables the user to easily find emotions related to the one in which s/he is interested. Grids whose co-occurrence frequency is lower than a threshold are displayed in a light color so that the user can focus on the related emotions, although the user can click a light-colored grid if s/he is interested in that emotion.

B. Search Result Interface

Area Ⓒ in Figure 3 is the search result interface for the selected emotion. It consists of a ranked video list, video player, and line chart. We describe the details of each of these elements below.

[Ranked video list] When a user selects an emotion, SmartVideoRanking generates a list in which the top five videos are ranked in terms of intensity of the selected emotion. "Intensity" means the number of posted comments that correspond to the emotion. In this example, videos in which more viewers post the comment "cute" and similar comments are ranked higher. We describe how to rank videos in Section V-E1. For each video, SmartVideoRanking displays a thumbnail, title, number of views, favorites, and comments. This ranked list enables the user to browse various videos related to the emotion in which s/he is interested.

[Video player] When a user clicks a video's thumbnail in the ranked list, the video is displayed on the video player. By default, the top ranked video is displayed. This video player is provided by NicoNico, and it can be embedded by writing a designated tag in HTML. The user can make the player full-screen by clicking the full-screen icon in the player.

[Line chart] SmartVideoRanking also displays a line chart for the played video. The horizontal axis in the line chart represents the video's playback position and the vertical axis represents the number of comments corresponding to the selected emotion. This line chart enables the user to understand the change in intensity of the emotion in the video. Moreover, the user can click the dots in the line chart to move to the corresponding part of the video. This enables him/her to watch a scene where the selected emotion is especially intense by clicking the dots with a high vertical axis value. In this example, the user can watch an especially cute scene by clicking the dot Ⓓ in Figure 3.

IV. Comment Feature

We compute the usefulness of emotions for searching videos by using SVM regression. Here, a useful emotion is one that users want to use for searching for videos. In this section, we first describe the data set of the study. Next, we describe the

preprocessing of comments to compute their feature values. Then, we describe the target comment collections to estimate usefulness and present the features used for SVM regression. Finally, we describe the method to sort grids on SmartVideo-Ranking.

A. Data Set

We used data provided by the National Institute of Informatics[6]. This data includes video metadata and comment data for about 8.3 million videos uploaded to NicoNico before the beginning of November 2012. With respect to the video metadata, we used tags and the number of views, comments, and favorites, while for each comment, we used the comment's text, posted time, and playback position at which the comment was posted[7].

Among the 8.3 million videos, we used videos in the following five categories: (1) an original song created using VOCALOID, (2) cover of an original song, (3) dancing to an original song, (4) playing a musical instrument to an original song, and (5) creating a music video for an original song. We categorized videos on the basis of data provided by Hamasaki and Goto [27]. Videos with $\geq$ 1,000 comments were selected, and for each selected video, the 1,000 most recent comments were included in our data set. Our data set consisted of 11,180 videos and 11.18 million comments in all categories.

B. Comment Preprocessing

Brody and Diakopoulos [28] reported that comments on Twitter are often lengthened by repeating certain letters, as in "cooooooolllllll", to emphasize an emotion. This is also common on NicoNico. Therefore, there are many comments that are different in the number of repeated letters, even though they have the same meaning, such as "cuuuuttteee" and "cuuuuuute". If we treat these comments as different, their appearance frequency would be relatively low, resulting in a sparsity problem when computing the feature values. To solve this problem, we used the method proposed by Brody and Diakopoulos [28] for normalizing words and associating them with a canonical form. Yamamoto and Nakamura [17] also normalized comments on NicoNico by following Brody and Diakopoulos [28] when they estimated music mood categories. In our study, the comments were normalized as follows.

1) Remove symbols such as "!" and "?".
2) Convert lowercase characters to uppercase ones (*e.g.*, "great" is converted to "GREAT").
3) Replace repeated letters with a single letters (*e.g.*, "CUUUUTTTTTEEE" is converted to "CUTE").

The above operations normalize "cuuuuttteee!!!!" and "cuuuuuute" to "CUTE".

It is not always appropriate to show normalized comments to users because some normalized comments do not have the correct spellings, *e.g.*, "realy" instead of "really". To solve this problem, given a normalized comment c, we find the most common original comment c_t in terms of its appearance frequency in our data set and regard c_t as the canonical form of c. When c is estimated to be useful, c_t is displayed to the user.

[6]http://www.nii.ac.jp/dsc/idr/nico/nico.html

[7]The IDs of users who posted comments are not provided so as to maintain privacy.

C. Comment Collection

Given a query q, we regard videos annotated with a tag of q as relevant and refer to the set of such videos as V_q. To eliminate comments that are rare even after normalization, we collect target comments that are posted on ≥ 3 videos in V_q and whose appearance frequency is ≥ 10 in V_q and use them to estimate the usefulness. Hereafter, we will let C_q denote the set of target comments.

D. Feature

Below, we describe the features used for a normalized comment c that corresponds to c_t. Although two features (CHORUS and SCHORUS) can be used only for music videos, the 12 other features can be used for any kind of video.

[Query relevance (QREL)] We assume a comment that is highly relevant to a query is useful and use the expected mutual information [29] as a feature, which is given by

$$epmi(q, c_t) = P(q, c) \cdot log \frac{P(q, c)}{P(q)P(c)}, \quad (1)$$

where $P(q) = \frac{|V_q|}{|V|}$, $P(c) = \frac{|V_c|}{|V|}$, and $P(q, c) = \frac{|V_q \cap V_c|}{|V|}$. V represents all videos in our data set and V_c represents the set of videos to which c was posted in V.

[Query similarity (QSIM)] We assume the comment (*e.g.*, Miku cuuuute!), which is similar to the query (*e.g.*, Hatsune Miku) is useful and use a normalized Levenshtein distance metric [30] as a feature, which is expressed as

$$qsim(q, c_t) = 1 - \frac{D_L(q, c_t)}{L_{max}(q, c_t)}, \quad (2)$$

where $D_L(q, c_t)$ represents the Levenshtein distance and $L_{max}(q, c_t)$ represents the maximum number of letters of q and c_t.

[Number of letters (LEN)] Although useful words might be included in a long comment, in this study, we do not extract words from a comment to simplify the problem. (Extracting words from long comments would be an interesting future work.) In this case, long comments tend to be a spam, so we assume such comments are useless; short comments are useful. We thus use the reciprocal of the number of c_t's letters as a feature.

[Appearance frequency (FREQ)] We assume a comment input by many users is useful and use the appearance frequency of c in V_q as a feature.

[Original comment variation (VAR)] If there are many types of lengthened comments that are normalized to c in C_q, it means c is associated with subjectivity and sentiment [28]. We assume such a comment is useful and use the number of unique lengthened comments that are normalized to c in C_q as a feature.

[Adjective (ADJ) / Adjective verb (VADJ)] We assume a comment including an adjective/adjective verb is useful and use its existence or non-existence as a feature: 1 for existence and 0 for non-existence. The existence or non-existence of an adjective/adjective verb is judged using MeCab [31], a Japanese morphological analyzer. Yamamoto and Nakamura [17] reported that an adjective/adjective verb is an effective feature to predict music video mood categories.

[Entropy (ENT)] In NicoNico, similar comments are often posted at similar playback positions in many videos. For example, the comment "thank you for uploading the video" is often posted at the beginning of a video. We assume that such formal comments are useless, whereas comments that are independent of the playback position are useful because they would be posted in response to the video content. On the basis of this assumption, for each video in V_q, we divide the video length into 20 blocks and for each one create a histogram that represents the playback position and c's appearance frequency in each block; then we aggregate the histograms. We then compute the entropy of the aggregated histogram and use it as a feature.

[Probability of occurrence in chorus sections (CHORUS)] Yamamoto and Nakamura [17] reported that comments posted in chorus (a.k.a. refrain) sections were good features to predict music video mood categories. Therefore, we assume such comments are useful and compute the ratio of c posted in chorus sections in V_q to the total appearance frequency of c in V_q. The ratio is used as a feature. We use the method proposed by Goto [32] to detect the chorus sections of a music video.

[Number of unique similar comments (SNUM)] When a comment is similar to c, it represents a similar meaning to c. We assume that the more similar comments c has, the more important the matter referred to by c is; that is, c_t is useful. In this study, we empirically define a comment that has the same first two letters as c and has ≤ 0.4 in terms of the normalized Levenshtein distance metric of c as a similar comment to c. Accordingly, the number of unique similar comments in V_q is used as a feature. Hereafter, let S_c denote the set containing c and its similar comments.

[Appearance frequency considering similar comments (SFREQ)] We compute the appearance frequency of the comments in S_c as described in FREQ and use the total appearance frequency as a feature.

[Entropy considering similar comments (SENT)] We create a histogram based on the comments in S_c as described in ENT and use the histogram's entropy as a feature.

[Probability of occurrence in chorus sections considering similar comments (SCHORUS)] We compute the probability of comment occurrence in chorus sections based on the comments in S_c as described in CHORUS and use the probability as a feature.

[Character bigram (BGR)] We assume letters included in a useful comment for a query are also useful for other queries and use the character bigram of c as a feature. Specifically, we extract character bigrams from all comments in our data set and use the existence or non-existence of each bigram in c: 1 for existence and 0 for non-existence. Note that other features have a one-dimensional value, while this feature is a multidimensional vector.

E. Sorting Grids

As mentioned in Section III-A, when a user clicks a grid, SmartVideoRanking sorts other grids on the basis of relatedness between comments. In this study, the relatedness between normalized comments c_1 and c_2 is computed from the co-occurrence frequency in the videos. Specifically, the degree of relatedness is given by $rel(c_1, c_2) = \frac{|V_{q,c_1} \cap V_{q,c_2}|}{\sqrt{|V_{q,c_1}||V_{q,c_2}|}}$, where V_{q,c_1} represents the set of videos to which at least one comment in S_{c_1} was posted in V_q. Grids whose relatedness value is less than 0.2 are displayed in a light color as mentioned in Section III-A.

V. EVALUATION

We attempted to answer the following three research questions through our evaluation: (1) how accurately can comment usefulness be estimated by using the proposed features? (Section V-C); (2) to what extent were detected emotions different from one query to another? (Section V-D); and (3) is a ranked video list based on an emotion different from one based on a conventional ranking metric? (Section V-E).

A. Query

We could not select queries from query logs for our evaluation because the query logs of NicoNico were not available. Carman *et al.* [33] reported that the vocabulary used for tags is similar to that of queries; therefore, we used tags as queries. Specifically, for all tags T in our data set, we counted the number of videos that were annotated with $t \in T$. Then, we selected the top 50 tags in terms of this number and used them in our evaluation. The maximum, minimum, mean, and median values of the received videos were 5,032, 117, 546, and 309, respectively.

B. Gold Standard

To evaluate the accuracy of the comment usefulness estimate, it is necessary to create a gold standard. However, the average number of normalized comments C_q for the queries used in this experiment was 2789.5, and manually labeling the usefulness scores of all of them would have been extremely time-consuming. Thus, we first divided comments in C_q into 20 groups according to the $epmi(q, c_t)$ score. For example, if C_q contained 1,000 comments, the comments between the 1st and 50th ranks in terms of $epmi(q, c_t)$ belonged to the first group, those between the 51st and 100th belonged to the second group, and so on. Then we randomly selected 10 comments from each group and obtained 200 sampled comments for each query.

To create a gold standard, we recruited eight assessors, all of whom routinely use NicoNico. Six were male and two female; their average age was 22.5 years, with a standard deviation of 3.7 years. In the assessment process, we first showed a query to the assessors and then showed them 200 sample comments,

TABLE I
SPEARMAN'S RANK CORRELATION, KENDALL'S τ, AND RMSE. SEE SECTION IV-D FOR DETAILS ON THE FEATURES IN THE "SELECTED FEATURE" COLUMN.

Step	Selected feature	Spearman	Kendall	RMSE
1	BGR	0.7454	0.6003	0.2779
2	LEN	0.7526	0.6075	0.2744
3	ADJ	0.7540	0.6086	0.2737
4	SFREQ	0.7546	0.6092	0.2737
5	VADJ	0.7547	0.6092	0.2736
6	QSIM	0.7547	0.6094	0.2735
7	SENT	0.7552	0.6099	0.2735
8	ENT	0.7555	0.6102	0.2735
9	CHORUS	0.7554	0.6100	0.2735
10	QREL	0.7556	0.6103	0.2735
11	SCHORUS	0.7555	0.6102	0.2735
12	VAR	0.7562	0.6109	0.2735
13	FREQ	0.7561	0.6109	0.2735
14	SNUM	0.7561	0.6109	0.2735

which were randomized for each assessor. When the query was "Dance" and the comment was "so cool," we asked "Do you want to search for videos on 'Dance' in which viewers say 'so cool'?". Assessors answered each question on a scale of 0 to 2, where 0 indicated "I do not want to search," 1 indicated "I somewhat want to search," and 2 indicated "I want to search." Each question was answered by five assessors. The average value of the five assessor answers was used as the gold standard of the comment's usefulness. In this way, we asked a total of 10,000 questions for 50 queries. We paid the assessors $2 for every 200 comments labeled.

C. Expecting Usefulness

1) Procedure: SVM regression was used on the gold standard data set in order to estimate the usefulness of each emotion (*i.e.*, a comment). We used LIBSVM [34] with standard parameterization to develop the model. In each feature, values were normalized to fit into the interval [0, 1] by min-max normalization. We used a leave-one-out cross validation over the 50 queries: 49 queries (*i.e.*, 9,800 comments) were training data and the remaining query (*i.e.*, 200 comments) was test data in each validation. To evaluate the SVM results, we used three evaluation metrics: Spearman's rank correlation, Kendall's τ, and root mean square error (RMSE). Spearman's rank correlation and Kendall's τ range from -1 (completely different rankings) to 1 (equal rankings). The RMSE is equal to or higher than 0, and a small value indicates high accuracy of the usefulness estimation.

We also evaluated important features using stepwise feature selection as follows:

1) Let the set of features in Section IV-D and the set of features used for SVM regression be F_{all} and F_{use}, respectively. F_{use} was initialized to ϕ.
2) Find the most effective feature $f \in F_{all}$ that minimizes the average RMSE over 50 queries by adding f to F_{use}.
3) Move f from F_{all} to F_{use} and go to (2) unless $F_{all} = \phi$.

2) Results: Table I shows the changes in the three evaluation metrics during feature selection. The values were averaged over 50 queries. The results show that the character bigram is quite effective for estimating comment usefulness. This means that a bigram in a useful comment for a query is also useful for other queries. The RMSE reached a minimum when BGR, LEN, ADJ, SFREQ, VADJ, and QSIM were combined, and no improvement was observed by adding other features. When the RMSE reached a minimum, Spearman's rank correlation was 0.7547, which indicates that the usefulness was accurately estimated.

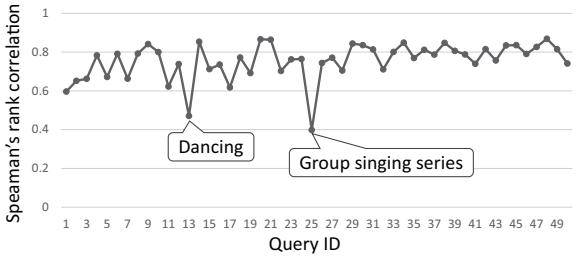

Fig. 4. Per-query Spearman's rank correlation.

TABLE II
COMPARISON OF ESTIMATED RESULTS AND GOLD STANDARD FOR QUERY "MUSIC". NUMBERS IN PARENTHESES INDICATE ANSWER VALUES.

Rank	Estimated results	Gold standard
1	Quite harmonious (1.6)	Really harmonious (2.0)
2	Pictures and lyric are good (1.6)	Headphones recommended (1.8)
3	Good pianist (1.8)	I was enthralled (1.8)
4	Really harmonious (2.0)	Good pianist (1.8)
5	I got goose bumps (1.2)	Mark this song as a favorite (1.8)
6	Great Len and Rin (1.0)	Great song (1.6)
7	OMG (1.4)	Good PV (1.6)
8	Funny (1.2)	I want this song's CD (1.6)
9	Rin is cute (1.0)	Congrats for million plays (1.6)
10	I was enthralled (1.8)	Quite harmonious (1.6)
11	Beautiful voice (1.2)	Pictures and lyrics are good (1.6)

Figure 4 shows the per-query Spearman's rank correlation for the feature set that reached the minimum RMSE. Although most query scores were higher than 0.6, those of the "dancing" and "group singing series" were low compared with other queries. One reason for the low scores is that comments containing user names are often posted on videos related to those queries. For example, on NicoNico, there are several videos in which a user "Mamu" dances. Viewers post comments such as "Mamu skillful" rather than just "skillful" to the videos. Because of the string "skillful", the comment "Mamu skillful" is also estimated to be useful by SVM regression; in fact, the estimated score of "Mamu skillful" was 1.342 and ranked sixth among the 200 comments. However, assessors judged that comments posted to a specific user were not useful, and the answer value was low: 0.4. There are Web pages listing the user names of singers and dancers who posted videos on NicoNico; one solution to this problem is to use the existence or non-existence of a user name in a comment as a feature based on those data.

Finally, Table II compares 11 comments with answer value $\geq$ 1.6 and the top 11 comments in terms of estimated usefulness for the query "music". Five comments among the top 11 comments were included in the gold standard. In addition, the scores of all comments in the top 11 estimated results were

TABLE III
NUMBER OF OVERLAPPING TOP k EMOTIONS OVER 50 QUERIES.

k	Average	Standard deviation	Min	Max
10	0.3527	0.6374	0	3
20	1.011	1.224	0	7
30	1.686	1.793	0	10
40	2.328	2.240	0	12
50	2.994	2.755	0	17

equal to or higher than 1.0. When comments are shown to users, it is crucial to eliminate useless comments; thus, the results for "music" were ideal for our research objective.

D. Emotion Overlap

This section answers the second research question: to what extent were detected emotions different from one query to another? If the detected emotions are largely different from one query to another, we can consider it meaningful to detect emotions according to the input query.

To answer the question, for each of the 50 queries, we collected the top k comments in terms of usefulness score estimated from the feature set that exhibited a minimum RMSE in the previous evaluation, *i.e.*, BGR, LEN, ADJ, SFREQ, VADJ, and QSIM. Then we computed the comment overlap between any two queries, which means a total of ${}_{50}C_2 = 1,225$ pairs of queries were computed.

Table III shows the results for k =10, 20, 30, 40, and 50. It is clear that the degree of emotion overlap between queries is quite low: less than three, even for top 50 emotions. Although a bigram is an effective feature to estimate the emotion usefulness, as mentioned in Section V-C2, the results in this experiment indicate that useful emotions are subtly different from one query to another, and it is therefore meaningful to detect emotions according to the input query.

E. Overlaps of Search Results

Our system generates a ranked video list based on an emotion selected by a user. If most videos in the ranked list overlap those from conventional ranking systems based on the number of views or comments, it is useless to generate a video ranking based on emotions. In this section, we discuss the evaluation of the overlap between videos in the ranked list based on an emotion and those in conventional ranking systems.

1) Procedure: Similar to Section V-D, we obtained the top 10 useful emotions for each query and evaluated the video ranking generated by q and c as follows:

1) Obtain videos in V_q to which at least one comment in S_c was posted and rank the videos in descending order of the appearance frequency of comments in S_c[8]. Let $R_{1,l}$ denote the set of top l videos.
2) Let $R_{2,l}$ denote the set of top l videos in V_q in terms of number of views.

[8]As mentioned in Section III-B, when a user clicks a grid, SmartVideoRanking generates the ranked video list for the selected comment in this manner. The thumbnail displayed in a grid is the thumbnail of the 1st ranked video.

TABLE IV
NUMBER OF OVERLAPPING TOP l VIDEOS OVER 50 QUERIES. NUMBERS IN PARENTHESES INDICATE STANDARD DEVIATIONS.

Ranking metric	Top 5	Top 10	Top 20	Top 30
# of views	0.04703 (0.2282)	0.3237 (0.6646)	1.186 (1.461)	2.813 (2.889)
# of favorites	0.08961 (0.2999)	0.3149 (0.6312)	1.299 (1.460)	2.896 (2.768)
# of comments	0.03870 (0.1931)	0.2594 (0.5591)	1.062 (1.399)	2.406 (2.612)

3) Count the number of overlapped videos between $R_{1,l}$ and $R_{2,l}$. In this evaluation, l was set to 5, 10, 20, and 30.

Similarly, we computed the overlap with respect to the number of favorites and comments.

2) Results: Table IV shows the number of overlapping videos in terms of number of views, favorites, and comments over 50 queries. In all metrics, the number of overlapping videos was low: less than three, even in the top 30. These results indicate that ranking videos on the basis of an emotion enables users to browse videos that do not appear at the top of conventional search results. This helps users to search for desired videos because most users are likely to look at only the top of the search results [35].

VI. USER STUDY

We conducted a user study to obtain user feedback on SmartVideoRanking.

A. Procedure

We recruited nine participants for our user study. Eight were male and one was female; eight were computer science students and one was a musician. The average age of the participants was 22.8 years, with a standard deviation of 3.3 years. Each participant was compensated with $10 in cash. All participants used a laptop running Windows 8.1 with a 1600 x 900 pixels display and a mouse. Four of them used their own earphones and five used earphones we provided.

At the beginning of the study, we introduced the participants to the concept of SmartVideoRanking and its use. After the participants became familiar with the system, we asked them to use it freely for 20 min. The participants' operations, such as clicking a line chart and playing a video, were recorded using screen recording software. Finally, the participants were asked to fill out a post-experiment questionnaire containing five questions on the usability of SmartVideoRanking using a 7-point Likert scale. They also provided feedback on the pros and cons of the user interface. Finally, we conducted an audio-recorded interview based on the questionnaire. The entire procedure for one participant took approximately one hour.

B. User Study Results

Table V summarizes the results of the post-experiment questionnaire, where the "Pos" column represents the ratio of positive answers ($\geq$ 5 on a 7-point Likert scale). As seen, the participants found SmartVideoRanking suitable for

TABLE V
RESULTS FROM POST-EXPERIMENT QUESTIONNAIRE.

Question	Mean	SD	Pos
1 I would like to use it frequently.	5.89	0.87	9/9
2 I found it unnecessarily complex.	1.44	0.50	0/9
3 I thought it was easy to use.	6.33	0.47	9/9
4 I needed technical support to use it.	2.33	1.15	1/9
5 I thought it is suitable for searching for videos.	5.78	1.03	8/9

searching videos (Q5), easy to use (Q3), and said that they would like to use the system frequently (Q1). One participant commented, "*It is interesting to search for a video based on how viewers respond to videos.*" Another participant noted, "*I was able to see many interesting comments that I would never have hit upon using just a search keyword.*" Three participants mentioned that they were originally not interested in VOCALOID videos, but they watched VOCALOID videos with interest using SmartVideoRanking. This indicates that video search based on viewer emotions has the ability to expand the users' interests. Although one participant marked 5 points for Q4 (I needed technical support to use it), the participant commented, "*It might be difficult to fully utilize the system without technical support such as a tutorial, but it was easy to use once I learned how.*" Below, we describe the user feedback regarding each interface and function.

[Grid style response interface] All participants appreciated the grid style interface. In terms of grid size, among all the initial clicks after selecting a query, 52.7% of them selected the largest or the second largest grid. Some participants said they selected the larger grid first because it was recommended by the system. These results demonstrate the positive effect of changing the grid size according to the usefulness of each comment.

In terms of the sorting function, only three participants actively clicked the grid that was placed near the selected grid. It turned out that there are two main reasons other participants did not think this function was useful. The first is that comments placed near the selected comment are not always related to it. This indicates that it is insufficient to compute the degree of relatedness between comments on the basis of their co-occurrence in videos. The second reason is that when comments with high similarity with the selected comment were placed near the selected comment, participants thought that it was pointless to click them. However, if we can display truly related comments that are not similar to the selected comment, sorting grids can be beneficial. It would be an interesting future work to compute the degree of relatedness more accurately by considering other factors such as the distribution of playback positions at which the comments were posted.

In terms of the thumbnail, six participants said they took the thumbnail into account to decide whether to click the grid. Most of them clicked the grid if the comment was interesting and the look of the thumbnail suited their preferences. Two participants said that when there was a gap between the comment and the thumbnail in a grid, they thought it was interesting and clicked the grid.

[Line chart] All participants remarked that displaying the line chart was beneficial for watching videos. Use of the line chart was classified into the following three categories:

- Seven participants clicked the line chart almost immediately after they played a video to move to the scene with the largest number of selected comments. If they were interested in the scene, they watched the video from the beginning; otherwise, they stopped watching it. This means that the line chart can be used as a clue to determine whether a user will watch the video. One participant commented, "*When I want to search for many interesting videos, the line chart is useful because I can effectively check many videos in a short time.*"
- One participant clicked the line chart after he finished watching the video. The participant mentioned that he used the line chart to watch scenes related to the selected comment again. This indicates that a user can use the line chart to review the video.
- One participant did not click the line chart during the 20 min of use. However, this does not mean that the line chart was not useful. The participant said that he watched both a video and its line chart in order to understand when the selected comment appeared. That is, the line chart is also useful to watch a video for predicting scene development.

Some participants suggested an improvement to the line chart. One noted, "*I want to see which part on the line chart is played on the video player so that I can understand how long it takes to achieve the peak of the line chart. Why not put a bar that moves from left to right on the line chart by synchronizing to the playback time?*" In the future, we plan to implement this function to make the line chart more useful.

[Ranked video list] Most participants noted that they watched videos below the second one in the ranked list if they were interested in the top-ranked video. SmartVideoRanking displays the top 5 videos, but four participants mentioned they wanted the system to display more. Moreover, although our system displays information such as a thumbnail and a title for each video, one participant commented, "*I want to know what comments are posted to each video in addition to the selected comment.*" We can realize this function by verifying that each of the top 50 comments for the query are posted to each video in the ranked list. One participant mentioned the synergistic effect of the grid display — namely, that a user can search for videos that are related to various kinds of comment by utilizing the grid display while s/he can also search for various kinds of video related to a specific comment by utilizing the ranked video list.

VII. CONCLUSION

We described SmartVideoRanking, a system that enables users to search for videos on the basis of viewers' emotions. SmartVideoRanking extracts viewers' emotions from time-synchronized comments and estimates the usefulness of each

comment by using SVM regression. Our main findings from the experimental results are:

- In terms of the estimation of comment usefulness using SVM regression, the RMSE achieved the minimum value when we used the bigram, the comment length, the existence of an adjective, the total frequency of similar comments, the existence of an adjective verb, and query similarity as the features. The resulting Spearman's rank correlation was 0.7547. The results indicate that SmartVideoRanking can display useful comments with high accuracy.
- In terms of the emotion overlap between queries, the number of overlapping emotions was less than three for the top 50 emotions. This result indicates that emotions for videos are different from one query to another and it is therefore meaningful to detect emotions according to an input query.
- Regarding the search result overlap between ranked video lists generated by viewers' emotions and ones generated by conventional metrics such as the number of views, the number of overlapping videos was less than three in the top 30. This result indicates that ranking videos on the basis of emotions enables users to browse videos that do not appear at the top of conventional search results.
- Participants of our user study found SmartVideoRanking suitable for searching videos, easy to use, and said that they would like to use the system frequently.

In this study, we utilized time-synchronized comments on NicoNico. However, posting comments at arbitrary playback positions in a video or song is now becoming popular. On Twitter, users post comments while watching TV programs, and on SoundCloud[9], users can leave comments at arbitrary temporal playback positions in music by clicking its waveform while they are listening to it. In the future, we plan to devise a TV program search and music search based on viewers' or listeners' emotions by utilizing such time-synchronized comments.

ACKNOWLEDGEMENTS

This work was partly supported by MEXT KAKENHI Grant Number (15H06887) and CREST, JST.

REFERENCES

[1] M. P. Kato *et al.*, "Investigating users' query formulations for cognitive search intents," in *SIGIR*, 2014.
[2] A. S. Sharma and M. Elidrisi, "Classification of multi-media content (videos on youtube) using tags and focal points," 2008, unpublished manuscript, http://www-users.cs.umn.edu/~ankur/FinalReport_PR-1.pdf.
[3] M. Cooper, "Presentation video retrieval using automatically recovered slide and spoken text," in *SPIE*, 2013.
[4] A. Ekin and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.
[5] M. Everingham, J. Sivic, and A. Zisserman, ""hello! my name is... buffy"–automatic naming of characters in tv video." in *BMVC*, 2006.
[6] I. Laptev *et al.*, "Learning realistic human actions from movies," in *CVPR*, 2008.
[7] K. Lee and D. P. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.
[8] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *CVPR*, 2009.
[9] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," in *MM*, 2000.
[10] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?" in *CSCW*, 2010.
[11] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Story-based video retrieval in tv series using plot synopses," in *ICMR*, 2014.
[12] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in *MM*, 2002.
[13] H.-B. Kang, "Affective content detection using hmms," in *MM*, 2003.
[14] M. Xu *et al.*, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.
[15] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *AAAI*, 2014.
[16] L. Pang and C.-W. Ngo, "Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos," in *ICMR*, 2015.
[17] T. Yamamoto and S. Nakamura, "Leveraging viewer comments for mood classification of music video clips," in *SIGIR*, 2013.
[18] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *CHI*, 2010.
[19] C. Cayari, "The youtube effect: How youtube has provided new ways to consume, create, and share music," *International Journal of Education & the Arts*, vol. 12, pp. 1–28, 2011.
[20] M. Hamasaki, H. Takeda, and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: Case study of hatsune miku videos on nico nico douga," in *UXTV*, 2008.
[21] K. Yoshii and M. Goto, "Musiccommentator: Generating comments synchronized with musical audio signals by a joint probabilistic model of acoustic and textual features," in *ICEC*, 2007.
[22] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," in *INTERSPEECH*, 2007.
[23] C. Liang *et al.*, "Script-to-movie: a computational framework for story movie composition," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 401–414, 2013.
[24] G. Van Oorschot, M. Van Erp, and C. Dijkshoorn, "Automatic extraction of soccer game events from twitter," in *DeRiVE*, 2012.
[25] S. Poria *et al.*, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, no. C, pp. 104–116, 2015.
[26] S. Nakamura and K. Tanaka, "Video search by impression extracted from social annotation," in *WISE*, 2009.
[27] M. Hamasaki and M. Goto, "Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community," in *WikiSym*, 2013.
[28] S. Brody and N. Diakopoulos, "Cooooooooooooooollllllllll-lll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs," in *EMNLP*, 2011.
[29] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. USA: Addison-Wesley Publishing Company, 2009.
[30] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.
[31] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to japanese morphological analysis," in *EMNLP*, 2004.
[32] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.
[33] M. J. Carman *et al.*, "A statistical comparison of tag and query logs," in *SIGIR*, 2009.
[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[35] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *SIGIR*, 2005.

[9] http://soundcloud.com