Live Semantic Sport Highlight Detection Based on Analyzing Tweets of Twitter

Liang-Chi Hsieh*, Ching-Wei Lee*, Tzu-Hsuan Chiu*
*Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
viirya@gmail.com, {r96944033, d98944011}@csie.ntu.edu.tw

Winston Hsu*†
†Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
winston@csie.ntu.edu.tw

***Abstract*—Microblogging as a new form of communication on Internet, has attracted the attention from researchers recently. Relying the real-time and conversational properties of microblogging, its users update their statuses and share experience within their the social network. Those characteristics also make microblogging an important tool for users to share or discuss real world events such as earth quake or sport game. In this paper, we propose a novel and flexible solution to detect and recognize real-time events from sport games based on analyzing the messages posted on microblogging services. We take Twitter as the experiment platform and collect a large-scale dataset of Twitter messages that are called tweets for 18 prominent sport games covering four types of sports in 2011. We also collect corresponding sport videos for those games. The proposed solution applies moving-threshold burst detection on the volume of tweets to detect highlights in sport games. A tf-idf-based weighting method is applied on the tweets within detected highlights for semantic extraction. According to the experiments we perform on the tweet and video datasets, we find that the proposed methods can achieve competent performance in sport event detection and recognition. Besides, our method can find non pre-defined tidbits that are difficult to detect in previous works.**

***Keywords*-Microblogging; event detection; burst detection**

I. Introduction

Microblogging as a new form of communication on Internet, has attracted much attention with its tremendous rise of popularity in recent years. By using brief text message (usually not longer than 150 characters), the users of microblogging services can update statuses, share experience and interests with their network of friends. Compared to traditional blogs that typically get updated by bloggers at intervals of a few days, microblogging users usually upload new messages more frequently and instantly. This real-time characteristic makes microblogging an important tool for users to share or discuss ongoing events in real world, such as earth quake or sport game. For example, the following are some posts extracted from an online microblogging service: *"it starts raining....", "yessirrrrr homerun andrew jones.. 2-1 #yankees"* and *"Congrats to the Dallas Mavericks - NBA World Champions. Enjoy the ring and being a champion. You've earned it!"*. Those are typical messages posted by users after they experienced or witnessed the events happened in real world and they wanted to share the experience immediately.

Because of the real-time characteristic of microblogging, researchers found that some important and interesting information and knowledge can be mined by analyzing the messages posted by users that are related to real world events such as election [1]. Besides, the users of microblogging services are considered by researchers as social sensors to detect real world events such as earth quake [2]. Recently, some researchers turned to use Twitter, one of the most notable microblogging services, to preliminarily support

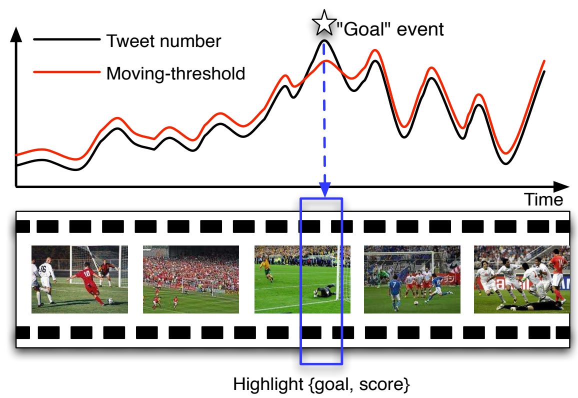


Figure 1. The illustration of the concept of live semantic sport highlight detection based on analyzing tweets of Twitter. By applying the moving-threshold burst detection proposed in this paper, we can find the highlights in the sport videos. Those highlights cover typical sport events such as "goal" event and some tidbits happened during the playing of the sport games. Non pre-defined tidbits are traditionally difficult to detect and recognize by previous event detection methods.

the event detection and recognition for sport games [3]. Compared to traditional event detection and recognition for sport games by visual content-based approach, the research direction demonstrated by [3], which relies on user reporting on Twitter to detect and recognize events in sport games, requires less computation and is not influenced by the visual quality of game videos. Because the microblogging-based approach does not train event models based on visual features of specific sport games but considers only the user activities on microblogging services corresponding to game events, this approach can be easier adapted to different types of sport games.

In this paper, we extend the microblogging-based approach of event detection and recognition for sport games as illustrated in Figure 1. We propose a more flexible solution for analyzing the messages shared between users on microblogging services to find important and meaningful events in sport games. This solution has two stages. First, for event detection, we propose moving-threshold burst detection that identifies the bursts in the volume of user messages on microblogging services corresponding to the happenings of events in sport games that are important to the users. Those detected bursts in the volume of microblogging messages are called *highlights* in this work. Second, by analyzing the textual contents of user messages detected and recognized as game events, we use a simple tf-idf weighting method to extract semantic meanings of detected events.

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.135

In many different microblogging services, we choose Twitter as our experiment platform because of its worldwide population and open, well-defined API to fetch data. As the usage of the microblogging services such as Twitter becomes widespread and the availability of Internet-enabled mobile devices continues to grow, people are more like to share their experience and feeling during watching sport games, no matter they watch the games at home or game arena. This study is motivated by this trend and we also want to explore the possibility of achieving competent event detection and recognition for sport games only depending on analyzing microblogging messages.

In this work, we experiment our detection and recognition method on four types of mainstream sports to demonstrate the possibility of generalizing the microblogging-based method across different types of sport games. Every type of the four sports, which are basketball, baseball, tennis and soccer, has different levels of popularity in various countries or regions. For example, baseball is most popular in the United States and soccer is popular in Europe and South America. Combined with the worldwide reach of Twitter, we can evaluate the proposed method on different sport types without regional bias.

The key contributions of this work are summarized as follows:

- We propose a flexible solution of microblogging-based event detection and recognition for sport game that applies moving-threshold burst detection on short messages posted by microblogging users.
- By applying the proposed method on Twitter, one of the most popular microblogging services, we identify ongoing events of sport games in real-time and extract semantic meaning for identified events.
- The proposed method is evaluated on four types of mainstream sports, which are basketball, baseball, tennis and soccer, to demonstrate the possibility of generalizing the method across different types of sport games.

This paper is organized as follows: In the next section we introduce the previous works related to microblogging and microblogging-based event detection and recognition. We describe the proposed method for sport event detection and recognition based on microblogging in Section 3. In order to evaluate our method, we collect related tweets from Twitter and broadcasting videos for different sport games. In Section 4, we explain the process of collecting the data. Further, we describe our experiment design and show the results in Section 5. In Section 6, we provide discussion and possible future works. We give our conclusion for this work in Section 7.

II. Related Works

Researchers are interested in the intention behind the users of microblogging. For example, Java et al. [4] have investigated the topological and geographical structure of the social network in Twitter. They wanted to understand user intention and the reciprocity behavior among users. In the work, Java et al. pointed out there are four main types of user intention: daily chatter, conversation, sharing information and reporting news. Another Twitter related study by Kelly [5], which analyzed the structure of Twitter, also showed that pointless babble, conversation and news are the main topics of tweets.

As the widespread popularity of microblogging services such as Twitter that has millions of users worldwide that constantly share information and report real-time news online, those users can be considered as observers of real world events. In [2], Sakaki et al. modeled Twitter users as distributed social sensors and analyzed the real-time nature of Twitter. They applied probabilistic models to detect real-world events and estimate event location. Based on the study, they announced a warning and reporting system for earth quake prediction.

Xu et al. [6] used web-casting text to automatically detect the highlights of video sports. Their system recognizes the time of clock from game videos. The identified timestamp combined with the text extracted from web-casting web sites can be used to find the video segments corresponding to the events described in the web-casting text. Because the text of web-casting web sites usually provides very precise commentary on sport games, this approach could reach near perfect precision and recall rates in event detection. However, the approach relying on web-casting text can not discover undefined events that may draw people attention during sport games.

In their works [1] [7] [8], David A. Shamma et al. studied the relationship between Twitter and ongoing real-world events. They mainly focused on the 2008 presidential debates of United States. They investigated the practice of user tweets during live media events and analyzed the changes of topics in live events. They proposed two metrics Importance and Chatness that can be used to measure how conversational and how important some tweets are in a period of time during the media events. They also studied the reciprocity relation among Twitter users by using the network graphs formed by the usage of hash tags (#) and mentions (@) on Twitter.

Considering other type of real world events like sport games, we assume that the social sensors represented by microblogging users can also help us detect important and meaningful events during real-time game broadcasting. By analyzing some tweets related to the NBA All-Star game, we found that people sometimes are more like to discuss the tidbits during sport games such as celebrity showing up and the performance of singers than the game playing. Compared to the previous works of sport event detection that could only identify regular sport events, our proposed method that only relies on the analysis of microblogging messages can discover those tidbits that are probably difficult to pre-define before those events happen.

The work from Lanagan et al. [3] for sport event detection is the most similar work to this paper. They used the stream of tweets to detect sport events and summarize related keywords of those events. This work focused on rugby and soccer games and achieved the accuracy nearly equivalent to traditional audio-visual content analysis of sport event detection. We have the similar motivation with [3]: detecting and recognizing sport events meaningful to microblogging users by analyzing their conversations on microblogging services during sport game playing. The audio-visual content analysis for sport event detection and recognition is time-consuming and can not be easily adapted across different types of sports. Our approach that only analyzes the huge volume of messages of microblogging services provides a new direction to solve the problem of real-time event detection and recognition.

III. Live Semantic Sport Highlight Detection Based on Analyzing Tweets of Twitter

We analyze the volume of tweets on Twitter during sport games playing and broadcasting to identify possible highlights in the games. We design a flexible two-stage solution to detect and recognize sport highlights. In the first stage, we propose moving-threshold burst detection to detect bursts in the volume of tweets on Twitter. Those bursts can be considered as the points of highlights happening in the sport games. This method does not rely on a pre-defined threshold to find bursts in the volume of tweets. Instead, a strategy of using moving-threshold is applied to identify the bursts of highlights in sport games. After the highlights of games are detected, in the second stage, we perform a tf-idf based weighting method to extract semantic meaning for those highlights.

A. *Moving-threshold burst detection*

Previous burst detection methods usually define a threshold before detecting possible bursts. The threshold significantly affects the performance of burst detection. So how to choose the threshold becomes an important issue and in most of cases it has correlation to the data that those methods are applied on. In other words, a threshold works on one dataset may not be also workable on another dataset. For example, M.Vlachos et al. [9] have explored MSN search engine using some festival keywords such as ”Halloween” and ”Thanksgiving”. They detected the bursts of the keyword frequencies in long-term time-series data. However, for real-time event detection and recognition of sport games, it is difficult to define a suitable threshold in advance.

We propose moving-threshold burst detection that is a flexible solution to solve burst detection for real-time event detection. This approach is based on a moving-threshold that is obtained by computing the mean and standard deviation of the number of tweets during the playing of sport games. First we define a sliding window with the length w. For the length w, we find in our experiments that 30 seconds can work well. In the time sequence $(t_1, ...t_n)$ during the playing of sport games, we have a sequence of sliding windows $(W_1, ...W_n)$. The sliding window W_i at time t_i contains $N(W_i)$ tweets. For W_i, we compute the mean and the standard deviation of the numbers of tweets $(N(W_1), ...N(W_i))$ by tracing back the data from W_1 to W_i. Then we define the moving-threshold at time t_i as MT_i. We call it is a burst during the playing the sport games, when the number of tweets at time t_i, $N(W_i)$, is larger than the moving-threshold MT_i. The algorithm is as follows:

1) Define the sliding window with length w
2) At the time t_i, obtain the sliding windows (W_1,...W_i) and the corresponding numbers of tweets $N(W_1), ...N(W_i)$, and compute the values:

$$mean_i = mean(N(W_1), ...N(W_i))$$
$$std_i = std(N(W_1), ...N(W_i))$$
$$MT_i = \alpha * (mean_i + x * std_i)$$

3) Highlight detected in W_i if $N(W_i) > MT_i$

There are two parameters α and x. The parameter α is used to relax the condition of moving-threshold and is set between 0.7 and 1.0. For the parameter x, we set it empirically between 1.5 and 2.0. Figure 2 demonstrates the example of detecting highlights by applying moving-threshold burst detection on the tweets related to UEFA Champions League championship game and posted online during the game playing.

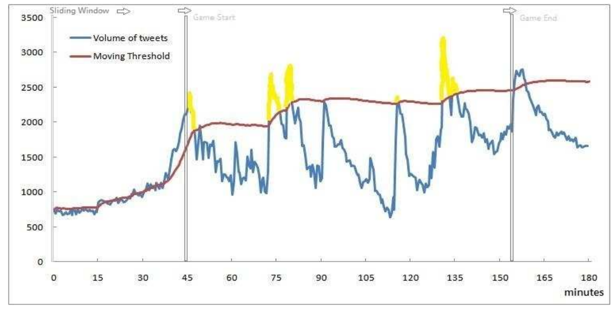


Figure 2. Detecting highlights by applying moving-threshold burst detection on UEFA Champions League championship game. Red line: the computed moving-threshold; Yellow lines: the detected highlights; Blue line: the distribution of number of tweets.

B. *Semantic extraction*

The highlights detected by applying moving-threshold burst detection represent the spikes of Twitter usage during the playing of sport games. In this second stage, we go further to extract the semantics of the detected highlights from the tweets of those highlights. We examine some example tweets as follows: *”gooooool de Barcelona”*, *”Barcelona 2:1 Man. United: GOAL: 54’Messi (Barcelona) scores! http://d3w.io/kRwtgV”*, *”#BAR Gol de Messi (min. 54) Barcelona 2-1 Manchester”*. We can construct a blurred scene of game playing by looking some keywords in those tweets such as ”goal”, ”Messi”, ”Barcelona”, ”Manchester” and the number of scores. The people who are familiar with sport game may know what those tweets talked is a soccer game. Actually, those tweets were posted by sport fans during the 2011 UEFA Champions League championship game, FC Barcelona vs. Manchester United. Those tweets are related to the highlight: at the first 54 minutes of the game, Lionel Messi from FC Barcelona scored a goal and made his team lead the game by 2-1. In this stage, we propose a method to extract such semantics from the tweets of highlights. This approach based on tf-idf weighting has the steps as follows:

Step 1:Apply standard stemming and stop-word elimination on whole tweet dataset.

Step 2:Calculate the idf scores of words in the tweet dataset for each sport domain, respectively.

Step 3:Aggregate the tweets for each detected highlight and calculate the tf scores of words within the time interval of each highlight.

Step 4:Rank the words for each highlight by using the tf-idf weight.

Step 5:Annotate each highlight with the higher ranked words, respectively.

The semantic extraction stage considers each tweet as a document and each detected highlight is a set of documents. In order to extract semantically significant words for each highlight, we apply standard tf-idf weighting method to rank the words in each highlight. In the information retrieval domain, tf-idf is widely used to measure the importance of a word. We measure the tf (term frequency) score of a word in a highlight as the number of times the word appears in the tweets of the highlight. And the idf (inverse document frequency) score of a word is measured by dividing the total number of tweets in the tweet dataset by the number of tweets containing the word in the tweet dataset and then taking logarithm of that quotient. It is noteworthy that a word might has different importance in different sport domains. For example, the word ”score” represents a significant event in soccer, but it has minor importance in basketball or tennis since there are usually numerous ”score” events happened in a basketball or tennis game. In order to obtain proper idf scores, we download tweets for each sport domain as training datasets before our evaluations to calculate the idf score for each sport domain.

IV. Data Collecting

We build our tweet dataset by collecting the tweets posted during several prominent sport games in 2011 from Twitter for four different sport domains: soccer, baseball, basketball and tennis.

For soccer, we choose UEFA (Union of European Football Associations) Champions League championship game between FC Barcelona and Manchester United. For basketball, we select 10 NBA (National Basketball Association) games, including the All-Star Game, 2 games from the Eastern Finals (Chicago Bulls vs. Miami Heat), 2 games from the Western Finals (Oklahoma City Thunders vs. Dallas Mavericks) and 5 games from the NBA Finals (Miami Heat vs. Dallas Mavericks). For tennis, Men's Championship games from the Australian Open, the French Open and the Wimbledon Open are included. Finally, for baseball, we choose 4 games from the regular season of MLB (Major League Baseball) between New York Yankees and Boston Red Sox.

A. Twitter dataset

As one of the most notable microblogging services, Twitter attracted more than 100 million new registered users and there were astonishing 25 billion tweets posted during 2010 alone [10]. How to download those data efficiently and automatically is a problem. Fortunately, Twitter provides high-throughput near-realtime access through its Twitter Streaming API for developers and researchers to download its data. Compared with the traditional Twitter API used by previous works to download tweets continuously with a constant time interval such as 30 seconds, Twitter Streaming API provides a better approach to collect tweets in a real-time and light-weight manner.

For the sport games mentioned above, we make real-time streaming connections with the API by keyword queries during the game playing. Those keywords include the game titles, the team names, the hashtags related to the game, the teams or some Twitter accounts such as the official Twitter accounts of the teams or fan clubs. Those query keywords we used are listed in Table II.

For every sport game, we do not only download the tweets posted during the game playing, but also the tweets posted before and after the game, in order to smooth the tweet distribution.

We have collected 617,950 tweets related to the soccer game, 1,455,151 tweets related to the basketball games, 169,892 tweets related to the tennis games, and 81,655 tweets related to the baseball games. Table I shows the details of the tweets. This large-scale tweet dataset includes more than two millions of tweets.

B. Video dataset

Besides the tweet dataset, we also collect the corresponding videos for those games. Those videos are used to examine manually the detected highlights. For the baseball and basketball games, we record the live streaming broadcasts from ESPN and STAR Sports TV channels. However, for the soccer and tennis games, we can not find any TV channel that broadcasts the games in our country. So we obtain the videos of the games via Internet from some users who have recorded those the broadcasts from ITV Sports and ESPN channels.

Table I
The statistics of the tweet dataset collected for 1 soccer, 10 basketball, 3 tennis and 4 baseball games, including the average tweets per minute and total tweets for each sport, respectively.

Games	Avg./Min.	Total tweets
Soccer	3,433.1	617,950
Basketball	808.4	1,455,151
Tennis	314.6	169,892
Baseball	85.1	81,655

Table II
Keywords used for Twitter Streaming API Query.

Sports	Keywords
Soccer	MUFootballClub, fcbarcelona
Basketball	NBA, allstar, okcthunder, chicagobull, miamiheat, dallasmav
Baseball	yankees, redsox
Tennis	ausopen, frenchopen, Wimbledon

C. Dataset alignment

Once those datasets are collected, we have found that the tweets extracted from Twitter do not synchronize well with the game videos in time. According to common sense, the tweets related to an event should be posted by Twitter users after the event happened. But in the collected datasets, we discovered that some tweets related to an event were posted before the event had happened in the sport game. The problem is caused by the satellite signal delay of streaming broadcasts. In order to synchronize the tweets and the events happened in the videos, we have to adjust the timeline for the collected datasets. However, the automatic alignment is difficult because of some reasons. Firstly, the satellite signal delay of live broadcasting stream is usually slight and acceptable to audience. So TV channels would not care about the delay too much and notice the delay. Secondly, although some websites provide web-casting text of sports games, the information is not always sufficient. For example, MLB Scoreboard of ESPN [11] provides the very rich data for baseball games, but the precise time information is lacking due to the nature of baseball game. Lastly, as the popularity of mobile devices, it is very likely that Twitter users post tweets from their mobile devices when they are watching a sport game at the scene. So those tweets are posted on Twitter much earlier than anyone who watches the corresponding broadcasting at home.

Because of these difficulties, we could not synchronize the tweets with the videos perfectly. Instead, we just slightly adjust the timelines of the tweets and the videos to eliminate unreasonable data such as that the tweets posted earlier than the occurrence of the corresponding events. For the videos recorded by other users, we had no clue about the timelines. Therefore, we only show the delay adjustment for the videos recorded by us in Table III.

Table III
SATELLITE SIGNAL DELAY ADJUSTMENT FOR THE VIDEOS (BASEBALL: 4, BASKETBALL: 10) RECORDED BY US.

Sports	Satellite signal delay adjustment (sec.)
Baseball	19.5, 13, 23, 18
Basketball	18, 20.5, 33, 19, 15.5, 17, 16, 14.5, 14, 22.5

Table IV
THE PRECISION OF DETECTED HIGHLIGHTS.

Sports	Soccer	Basketball	Baseball	Tennis
Precision	0.83	0.52	0.71	0.59

V. EXPERIMENTS

We apply the proposed method of sport highlight detection on those collected tweets and sport game videos in this section. First, we examine the correctness of highlights detected by moving-threshold burst detection. Then we inspect the precision of the results of the semantic extraction method. In the end of this section, we compare the proposed method with traditional event detection.

A. Precision of highlights

The highlights found by moving-threshold burst detection only represent the spikes of the number of tweets posted on Twitter during the corresponding time slot. However, if those detected highlights do not show any meaning to people, we will consider those highlights as false positive instances. In order to measure the precision of detected highlights, we examine the tweets within the highlights to see if there is any meaningful event happened in corresponding game videos.

In the experiments, we apply two rules to measure the precision of highlights.

- Rule 1: if a tweet has been retweeted by sufficient quantity of users, we consider that there is a central topic among those users. We set the quantity to 1% of the number of tweets. But the rule 1 is not a strong rule because there are more than 40% of tweets on Twitter that are referred to pointless babbles [5]. So we introduce next rule 2.
- Rule 2: we manually observe the tweets of a highlight and the words extracted from the tweets using tf-idf weighting. If we could recognize a meaningful event, we consider the detected highlight as a positive result.

All of the regular sport event are considered as positive examples here. For example: goal, shot and corner in soccer; score, hit and homerun in baseball; dunk, 3-point and alley-oop in basketball; ace, double fault and rally in tennis. We show the precision of detected highlights for every sport domain in Table IV.

B. Precision of semantic extraction

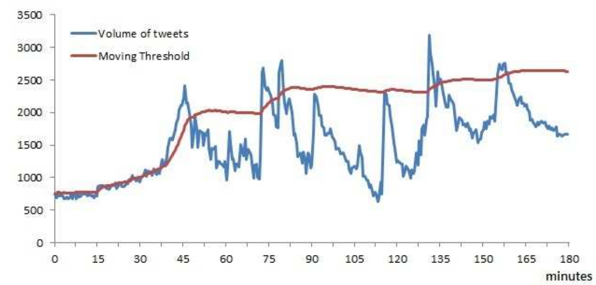
After we detected the highlights from the tweets, we apply the semantic extraction method on the tweets within detected highlights. The words used in the tweets within every highlight are ranked by tf-idf weighting. We show the average P@1, P@5 and P@10 performance of those extracted words for different sport domain in Table V.

C. Highlight detection for specific event

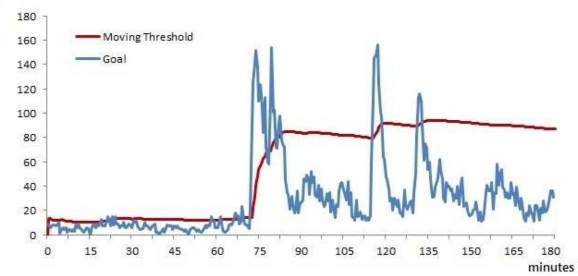
We have experimented the proposed highlight detection on all collected tweets. Further, we could also apply the moving-threshold burst detection on sampled tweets containing the keywords specifying special sport events. For every kind of event in sport, we sample the tweets from our Twitter dataset that contain targeting keywords of the event. Then we apply moving-threshold burst detection on the sampled tweets. This approach is similar to previous work of M. Vlachos et al. [9] who investigated festival events on MSN search engine by using keyword querying. We illustrate the distribution of tweets and the computed moving-threshold for 2011 UEFA Champions League championship game in Figure 3(a). Similarly, the distribution of tweets sampled by using keyword ”goal” and the computed moving-threshold for the same game are illustrated in Figure 3(b). Detailed performance in precision and recall of various sport events for different sport domain is shown in Table VI.

Table V
THE AVERAGE PRECISION OF TOP 1, 5, 10 WORDS, RANKED BY TF-IDF WEIGHTING, OF ALL DETECTED HIGHLIGHTS.

	Soccer	Basketball	Baseball	Tennis
P@1	1.00	0.49	0.72	0.53
P@5	1.00	0.33	0.58	0.42
P@10	0.43	0.28	0.33	0.29



(a) Distribution of all tweets



(b) Distribution of tweets containing the keyword ”goal”

Figure 3. Comparison of tweet distribution in time between all tweets in 2011 UEFA Champions League championship game and sampled tweets in the same game.

VI. DISCUSSION

In Table VI, we have shown the performance of specific event detection by apply moving-threshold burst detection on the tweets sampled with corresponding event keywords. First, it is observed that we can find all ”goal” events of the soccer game in the detected highlights. Because ”goal” event is the most important event for soccer game, the users of Twitter who watch the soccer game will post related tweets in very high probability when the kind of event

Table VI
THE PRECISION AND RECALL (P/R) OF THE EVENTS DETECTED BY MOVING-THRESHOLD BURST DETECTION FOR SOCCER 5(A), BASEBALL 5(B), BASKETBALL 5(C) AND TENNIS 5(D).

(a) Soccer events

Event	P/R	Event	P/R
Goal	1.00/1.00	Substitution	0.67/0.60
Shot	0.73/0.34	Yellow Card	1.00/0.25
Corner	1.00/0.67	Foul	0.75/0.19

(b) Baseball events

Event	P/R	Event	P/R
Score	0.77/0.43	Hit By Pitch	0.86/0.75
Hit	0.46/0.19	Ball Four	0.40/0.13
Homerun	1.00/0.67	Strike Out	0.00/0.00

(c) Basketball events

Event	P/R	Event	P/R
Dunk	0.43/0.08	3-point	0.53/0.03

(d) Tennis events

Event	P/R	Event	P/R
Ace	0.55/0.13	Rally	0.67/0.32

happens. However, compared to the results of basketball and tennis, the moving-threshold burst detection obtains poor performance in detecting the events such as "dunk" or "ace". We think that is because those events are occurred too often in the sports. So the Twitter users would be less interested in the occurrence of such events. The characteristics of those sport domains and their events can be used to explain the lower performance.

It is worth mentioning that in the 2011 NBA All-Star Game which was a grand party and competition, many celebrities showed up to watch the game or performed at the game. From the detected highlights of the game, we find some interesting tweets. For example, the tweet *"@LennyKravitz is killin the intros right now on @NBAonTNT http://twitpic.com/421vlx"* was posted when the singer Lenny Kravitz performed during the introduction of players before the game started. And another tweet *"RT @LordBieber: Sitten court side at allstar game with @justinbieber o yea what."* was posted when the singer Justin Bieber showed up in the stadium.

This findings show that people sometimes are likely to discuss such tidbits than typical events happened during a sport game. Our proposed highlight detection method based on the analysis of Twitter tweets can easily find those untypical events. Compared to our method, traditional event detection could only detect the typical sport events. The social interaction happened on the microblogging services such as Twitter opens a new direction for real-time event detection of real world events. There are many issues waiting for researchers to solve, such as the combination of content-based and microblogging service-based event detection methods. We leave this as our future works. The satellite signal delay is another issue in this study. But if we can obtain help from TV channels or the people at the scene of game playing, the data alignment would not be a difficult problem.

VII. CONCLUSION

In this work, we investigate the new direction to detect sport events by only using the social messages posted online. We propose moving-threshold burst detection to detect highlights happened during the playing of sport games. The flexible method will adjust the threshold for detecting bursts of the number of tweets. Then we apply tf-idf weighting method to extract the semantic meanings for the detected highlights. We collect a large-scale tweet dataset from Twitter and a sport video dataset for four different sport domains: soccer, baseball, basketball and tennis. The proposed methods are performed on the datasets to detect sport highlights with semantic meanings. From the experiment results, we find that the proposed methods can efficiently and effectively find typical sport events such as "goal" event. Furthermore, the proposed methods can also detect some tidbits such as a singer showing up in a sport game. The detection of such untypical sport events is less focused before in previous works. We think that this study has opened a new direction for real-time event detection of real world event.

REFERENCES

[1] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Proceedings of the first SIGMM workshop on Social media*, New York, NY, USA, 2009, WSM '09, pp. 3–10, ACM.

[2] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, WWW '10, pp. 851–860, ACM.

[3] James Lanagan and Alan F Smeaton, "Using twitter to detect and tag important events in live sports," *Artificial Intelligence*, pp. 542–545, 2011.

[4] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, "Why we twitter: understanding microblogging usage and communities," in *WebKDD/SNA-KDD '07*, pp. 56–65.

[5] "Twitter study reveals interesting results about usage," Tech. Rep., Pear Analytics, San Antonio, Texas, http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

[6] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan, "Live sports event detection based on broadcast video and web-casting text," in *MULTIMEDIA '06*, pp. 221–230.

[7] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events?," *CSCW Horizons*, 2010.

[8] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill, "Conversational Shadows: Describing Live Media Events Using Short Messages," in *4th AAAI Conference on Weblogs and Social Media*, 2010.

[9] Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos, "Identifying similarities, periodicities and bursts for online search queries," in *SIGMOD '04*, pp. 131–142.

[10] "Twitter's population growth: Where is everyone coming from?," *http://blog.hubspot.com/blog/tabid/6307/bid/8642/Twitter-s-Population-Growth-Where-is-everyone-coming-from.aspx*.

[11] "Espn mlb scoreboard," *http://scores.espn.go.com/mlb/scoreboard*.