Timelines as Summaries of Popular Scheduled Events

Omar Alonso
Microsoft Corp.
omalonso@microsoft.com

Kyle Shiells
Microsoft Corp.
kshiells@microsoft.com

ABSTRACT

Known events that are scheduled in advance, such as popular sports games, usually get a lot of attention from the public. Communications media like TV, radio, and newspapers will report the salient aspects of such events live or post-hoc for general consumption. However, certain actions, facts, and opinions would likely be omitted from those objective summaries. Our approach is to construct a particular game's timeline in such a way that it can be used as a quick summary of the main events that happened along with popular subjective and opinionated items that the public inject. Peaks in the volume of posts discussing the event reflect both objectively recognizable events in the game -- in the sports example, a change in score -- and subjective events such as a referee making a call fans disagree with. In this work, we introduce a novel timeline design that captures a more complete story of the event by placing the volume of Twitter posts alongside keywords that are driving the additional traffic. We demonstrate our approach using events of major international social impact from the World Cup 2010 and evaluate against professional liveblog coverage of the same events.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords

Timelines, sports, summaries, World Cup, soccer, football, rugby.

1. INTRODUCTION

Microblogging sites like Twitter have gained tremendous attention as platforms to disseminate information about ongoing events. One of the major uses of Twitter is to discuss events in real time. As users join a conversation with their friends and the Twitter community, they generate huge amounts of time-stamped text narrating and discussing the ongoing event. The character limit and time-sensitivity of the content lead people to publish so quickly that only live blogging, a curated form of microblogging, can allow the professional media to keep pace with Twitter in instantaneous coverage.

In the particular case of sports events, users tweet to express their opinions, joys, and frustrations as the game progresses, as well as to report the objectively observable events of the game. The aggregation of the subjective content seems useful as one can see microblogging as another channel to produce and consume polarized and personal information, in contrast to what one can expect from a professional transmission via TV, radio, or even a live blog. The advantages of scheduled events for the analysis we present are that an audience is established and prepared for discussion from the very beginning of the event and that the start and end times are clearly defined. Media game reports tend to be very objective and highlight certain aspects of the contest but usually miss other facts that a news consumer might also like to read about. Producing a compact summary of a sports game using Twitter data that not only covers the obvious events but also augments them with the crowd's reactions may provide a more complete and satisfying way to represent the full narrative of a game to users.

One can think of our approach as something similar to a *global bar atmosphere*, where fans from both teams playing are carrying on conversations simultaneously alongside other sports enthusiasts, who may not be invested in the specific teams playing but are there to enjoy the sport for its own sake. All the members of the crowd watch the game together and react immediately. Fluctuations in the volume of Twitter traffic can be thought of as differences in the audible volume of conversation in a bar. When a goal is scored, many people react at once and this combines into a roar that can confirm, even to an uninformed observer, that something important must have just happened. The rules of the game define the most salient events (e.g. goal, full-time) and fans will echo those outcomes along with subjective reactions to the events. In addition, fans might respond to a particularly clever play, or (as we will show in an example later) may respond out of proportion to one goal because it represents a milestone for the player who scored it – an event that would likely be mentioned by a newscaster, but requires contextual knowledge beyond the game itself. How can we capture both the objective and subjective elements of a game and combine them in a visualization of how that game unfolded over time?

Our method is to mine the tweets produced in the online conversation around the event and provide an aggregated summary of the game along a well-defined timeline, with the aim of capturing the most salient parts of the virtual bar activity. Note that we are not interested in sentiment analysis and do not use formal approaches to event or topic detection. Rather, our focus is on a precise timeline that presents the well known facts annotated with the most prominent keywords used by the crowd at a given time, where events can be observed as peaks with associated keywords but are not otherwise separate from the ongoing discussion. We believe this offers an interesting middle ground between the narrative continuity of an editorial summary, as would be presented by a TV broadcast or newspaper, and the huge volume of diverse text that various members of the crowd share in microblogs.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05.

We selected the FIFA World Cup 2010 as motivating scenario. Twitter reported record traffic during the WC matches[1] and was used widely around the world to discuss the games as they took place regardless of the time zone and geographical location. Because some teams have more active Twitter fans than others, we can show how our timeline summarization method reflects the composition of the virtual bar, whether both teams have vocal supporters, one team is much better represented, or the crowd in the bar is largely indifferent to both teams and simply there to enjoy the game.

There are a number of research questions that we would like to pose and will explore in detail with our study:

- Sports in general are well-covered by TV and radio by continuously describing the action. In absence of such media, would it be possible to follow a game via a microblogging site? We are not interested in replacing these media, but in comparing our approach to theirs in terms of informativeness.
- The outcome of a game is usually summarized by raw numbers. That is, number of goals, points, touchdowns, etc. depending on the sport. While these aggregate numbers represent the result, they fail to capture the rhythm and intensity. Can we present a game summary in a way that combines the main events along with salient comments and the dynamics of the game in a timeline?
- Would it be possible to construct a rich summary from the Twitter traffic around a sports event that contains aspects not covered by a traditional news article?

In this paper we make the following contributions:

- A new timeline display designed for sports that allow observers to compare games within the same sport and to a lesser extent across sports.
- Annotation of the timeline with popular terms that reflect what the crowd was discussing at key moments.
- A study of a few games of the World Cup using our techniques and some statistical insight into the dataset.

This paper is organized as follows. In the next section we cover related work on this area. In Section 3, we describe the World Cup data set and present some statistics. In Section 4, we introduce the timeline display and perform analysis on some World Cup games. The evaluation is presented in Section 5. In Section 6, we generalize the timeline generation to other sports events. Finally, we present conclusions and future work in Section 7.

2. RELATED WORK

Research work on Twitter data is very popular nowadays with the bulk of the work concentrating on topic detection and, to a lesser extent, sentiment analysis. In the last year or so, there has been exciting work on analyzing Twitter content in the context of major events. A clear example is the study by Gaffney [6] of the Iran elections. A study on users' responses to events involving actors and celebrities is presented in [14]. Event summarization using tweets in the context of football is studied by Chakrabarti and Punera [4]. Leveraging humans as real-time sensors with some examples from sport events including the World Cup is presented in [8] and [23]. New work that goes beyond trending topics with a focus on real-world events is described in [3].

The closest prior research to our work is TwitInfo, a system for visualizing and summarizing events [12]. The main differences with TwitInfo are that we concentrate on the timeline generation by presenting more variables on the display and providing a better baseline for comparison. A stream summarization approach using a small data set from the World Cup is presented in [19].

Multimedia, particularly video understanding, has been an active area of research aimed towards extracting good summaries of games. Examples of techniques for browsing sports videos and summarization of key events and players can be found in [9] and [20]. Some researchers are studying the effect of combining both video and microblogging streams such as [7], [17] and [21]. We focus only on tweets.

Research work on fully utilizing the temporal information embedded in the text of documents for exploration and search purposes is very recent. The NEAT prototype uses the NYTimes data set to populate a timeline using crowdsourcing for annotating the most important events. A similar approach to anchoring articles in time is given in [13]. An in-depth study of presenting past information using timelines is presented in [22].

There has been seminal work on temporal summaries of news topics by Allan, showing how important temporal information is [1]. Time information is also used in temporal mining of blogs to extract useful information [15] and temporal patterns [16] among other things. Extensions to document operations such as comparing the temporal similarity of two documents in the context of news articles are presented by Makkonen and Ahonen-Myka [10]. TimeMine combines topic detection and tracking with timelines as a browsing interface, presented in [18]. A more sophisticated visualization is described in [5]. Timelines as a mechanism for organizing content have received a lot of attention lately with the announcement of Facebook's timeline.

3. THE WORLD CUP DATASET

Our dataset from the FIFA World Cup 2010 (WC) consists of tweets containing the hashtag `#worldcup` from 12:00 June 5 to 12:00 July 17 2010. All times throughout this paper will be given in UTC, and the timestamps of all tweets were normalized to UTC before any processing. The tweets we collected arrived at an average rate of 63 per minute for the duration of the WC, out of a total average tweet rate of 22,136 per minute for the same time period. The number of total tweets in our data set is 2,901,840, written in over 50 languages. One can think of this data set as a *representative sample* of all tweets about the World Cup. The time period during which these tweets were collected was unprecedented in terms of Twitter traffic, and as a result there are some gaps in our data due to server outages and such. Between our hashtag filter and the data collection problems, we don't claim to have every tweet available, but we do still have a very large corpus and observe interesting trends from mining it.

A total of 32 teams representing their respective countries participated in the World Cup finals. There are different ways to mention a team in Twitter so in our research we use the hashtags

[1] http://blog.twitter.com/2010/07/2010-world-cup-global-conversation.html

used on the official FIFA World Cup website and in the BBC coverage, among other places.

We started collecting data a week before the first game and stopped a week after the final. As expected the distribution is mostly dominated by countries with an old tradition in soccer and high representation on Twitter. Table 1 shows the top-20 English keywords for the WC data set. We discarded stop words, hashtags, and keywords that reference countries and teams. Note the high frequency of "vuvuzela" (the famous plastic horn). It is also interesting to see that the keyword "soccer" has more frequency than "football," again probably due to the high number of US natives on Twitter.

Keyword	Frequency	Keyword	Frequency
Watch	306,880	Team	86,294
World	301,856	Today	77,837
Cup	284,089	South	74,069
Free	209,129	Final	74,063
Fifa	190,417	Live	72,098
game	182,824	Good	70,405
vuvuzela	178,453	soccer	69,519
match	159,631	:)	57,968
online	131,528	football	55,893
2010	127,268	Great	43,380
Twitter	101,413	Best	38,894

Table 1. Top-20 keyword distributions.

4. FRAMEWORK

4.1 Definitions

As the basis for anchoring tweets in time, we assume a discrete representation of time based on the Gregorian Calendar, with a *minute* being our unit of time. Our base timeline, denoted *Tg*, is an interval of consecutive minutes grouped in three segments that denote a short period *before* the game starts, the events *during* the game, and a short period *after* the game has finished. A key aspect of our approach is the use of the timestamp of a tweet to anchor the content in the *Tg*.

An advantage of scheduled events for our approach is that we know the exact start and end times. Even if the event runs long, there remains a distinct end. The temporal boundaries and the defining hashtag simplify the detection of the bulk of the activity that we are trying to extract. The huge volume of total Twitter traffic also means that we can construct useful aggregates without seeking to retrieve every single tweet about our targeted event. In our context, we define an event *e*, as an occurrence of importance defined by the mechanics of the game. In the case of soccer, examples of events are: goal, penalty, yellow card, red card, substitute, first half, second half, and full-time to name a few. There many different events within a game, defined at various levels of detail, so we only consider the most important events that a casual observer should be familiar with. While we expect more minor events to be discoverable from the commentary, we explicitly place goals, starts and ends of play, and cards in our visualization.

4.2 Display Design

The game timeline design consists of an x-axis with four clear black marks denoting the start and end of the first and second halves. On the same axis we mark cards (yellow and red) in their respective colors. On the very top we mark goals using two colors (green and light blue). Each color represents a team and they are the same independent of which teams are playing. We use corresponding colors to represent the tweet activity explicitly mentioning each team. A gray color is used to represent the total including others who may not use any team hashtag, or in fact any term we explicitly track besides the hashtag defining the dataset, in their tweets. For each peak, we present a frequent keyword, to show the focus of the discussion that drove the peak. The y-axis represents the overall tweet count.

The display allows a user to get a high-level overview of the game: the final score, which team scored first, key points of conversation during the game, whether the game was very physical (i.e. many cards were given), etc. The display encodes a number of variables: the temporal sequence of the game, the rules and outcome of the game, the total numbers, the activity per team, and the figurative noise level of the virtual bar. An advantage of our design is that it allows comparison between games across many dimensions, as we will show in the evaluation section.

4.3 Timeline Generation

To generate the timeline, we collect all of the tweets containing a hashtag connected with the central event, in this case the World Cup as a whole, within a range around the scheduled time of the game. We then group tweets into 5-minute time buckets. Within each bucket, we collect total volume counts as well as counts for each word. For the line portion of the timeline, we graph the volumes of a few hand-selected key words associated with the game. For the World Cup, these are "goal" and the shortened forms of the names of each team playing, e.g. `NED` and `URU` for the Netherlands vs. Uruguay match. Graphing these term frequencies over time provides insight into the flow of the game. In particular, it shows major spikes of "goal" when goals are scored, almost always accompanied by a larger spike for the name of the scoring team and a smaller spike for the name of the team scored against. At the start of half time and the end of the game spikes for the winning team are also apparent. When a first or decisive goal is scored, all these peaks tend to be much larger.

In addition to tracking the volumes of those few domain specific terms, we examine the rest of the counts to find which key words are driving the spikes. These are shown as labels on the timeline, particularly at spikes. We identify these terms by treating the bucketed counts as histograms then, based on the histograms, assigning z-scores to the volume of a term in any 5-minute interval as follows:

$$z_{w,t} = \frac{n_{w,t} - \bar{n}_w}{\sigma_{n_w}}$$

where $n_{w,t}$ is the count of term *w* within time interval *t*, $\bar{n}_w$ is the mean count of term *w* over all time intervals, and σ_{n_w} is the standard deviation of the counts of term *w* over all time intervals.

The terms with the highest z-scores for a given time interval are displayed on the graph alongside the line indicating total volume, and provide an explanation for each spike. In addition, when an objectively observable event such as a goal or yellow card occurs, this often shows what the viewer reaction was. For example, in

the match Slovenia vs. USA, one goal led (English-speaking) fans to respond with "crap," while another was labeled "equalizer."

Finally, in order to make the referents of some of these terms clear, we annotate the timeline with some objectively observable events. At the top of each timeline we show the goals, color-coded by the scoring team. At the bottom, we show the starts and ends of play and yellow cards. These are all drawn manually from a BBC live blog of the same event. While there are strong signals for the events in the term volume, we add these to show how the timing of the reactions relates to the actual timing of the events. Delays between events and reactions appear to vary partly because of the way our data is discretized. Because we treat 15:24:59 still as part of the 15:20 bucket, the reactions to a goal scored then will tend to appear in the 15:25 bucket.

Separately from the timeline, we produced sets of example tweets for each game that represent typical tweets during spikes. We do this by summing the z-scores of the "spiking" terms present in each tweet. Because of the presence of more spiking words, these typical tweets tend to cluster around peaks in overall traffic. This, however, only means that the more exciting moments of the game are better represented in our summary: a desirable property. Some examples of these tweets, filtered to illustrate the meanings of spiking terms shown in the charts, are presented in the tables accompanying the timelines in the next section.

4.4 Analysis of Popular Games

In the following we analyze a total of six games. Unfortunately, the data for the final game is incomplete due to an internal glitch so we were unable to produce such timeline. Nevertheless, the games selected represent some of the best in tournament: #22 (Slovenia-USA), #41 (Slovakia-Italy), #51 (England-Germany), #61 (Uruguay-Netherlands), #62 (Germany-Spain), and #63 (Germany-Uruguay). For each game we present the timeline and a few tweets about the keywords.

In Figure 1 we present the timeline of a major upset game. Note the huge spikes at the end of game due to the goals and also the comments about the elimination of Italy ("goodbye", "eliminated"). Despite the excitement of the game, however, neither of the countries playing is particularly well-represented on Twitter, so the reporting is fairly objective. At the beginning, the audience simply reports the kick-off, and similarly "*tiempo*" announces the beginning of half time and "2nd" the beginning of the second half. When the first goal is scored, the name of the scoring player is the leading comment.

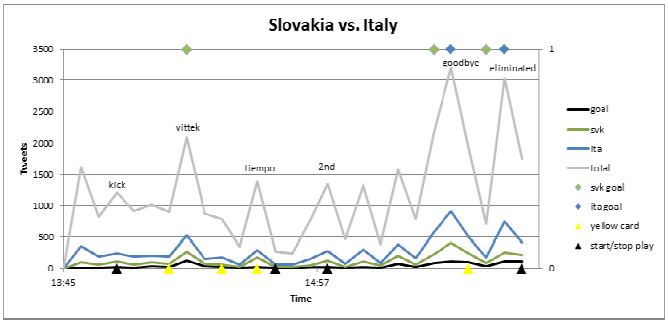

Figure 1. Game #41: Slovakia vs. Italy. Lots of activity at the end as it was a closed match that resulted in a huge upset.

```
What a game!!! amazing !!! Italy #ita 2 - 3 #svk
Slovakia Wow..... Bye Italy!! (What a shame) What
a big loser #worldcup winner (oops sorry)
Arrividerchi Italiano!! Arrividerci Cannavaro!!
Slovakia deserve it 2 win!! Good game, the best
game so far!! #worldcup
```

One of the most anticipated games (#51) is presented in Figure 2. The traditional anthem ritual is very noticeable and is, as indicated, a few minutes before the kick-off. The South African vuvuzelas are in full swing as the ball starts rolling. Note the presence of just numbers on the peaks and overall disappointment with the English team as the game progresses. In the case of "12", the reference is to Klose's 12th career goal in the world cup, tying Pele's record.

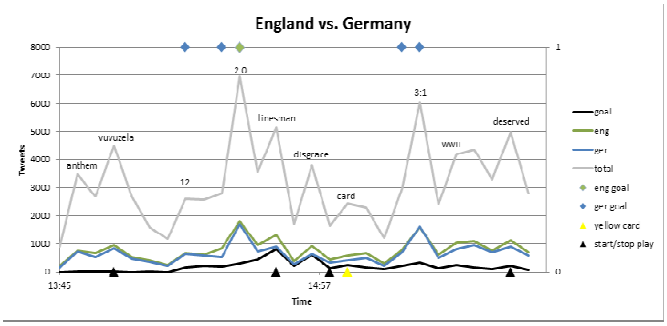

Figure 2. Game #51: England vs. Germany. Mostly English fans complaining about their team's performance.

```
Wow England got robbed from a clear goal!!! the
quality of the referees and lineman is ridiculous
this #worldcup
Oh crap. #GER scores on a counter attack. #ENG
couldn't run back in time. :( #worldcup
```

Game #41 is presented in Figure 3. The dip around half-time is due to missing data for a few minutes. Because of ongoing problems with data loss over the course of the game, evident to a lesser extent in the second half, the volume line in this graph is not as reliable. We still include it, however, because it strongly contrasts the objectivity of Figure 1. Since one team's (USA's) fans are much more prolific on Twitter, more sentiment comes across in the extracted keywords. When Slovenia scores their first goal, the fans are disheartened, when USA scores their first the fans are relieved, and during a lull in the action they produce text art to cheer their team on.

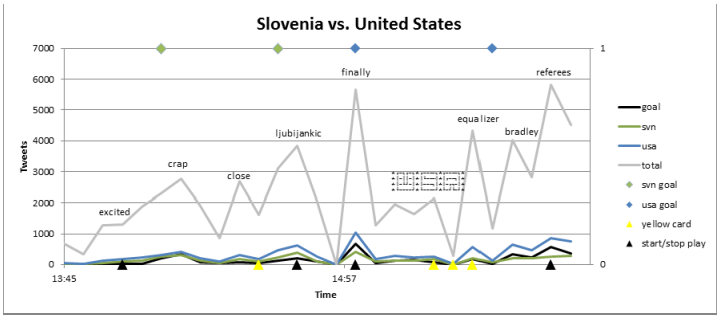

Figure 3. Game #22: Slovenia vs. USA. High activity in the second half as USA tied in the last few minutes.

```
#USA should have won that match. The refs took
that one away! Either way a great comeback. Great
hard earned point! #WorldCup
this game killed me. im glad we came back. and
tied this game after it looked absolutely
hopeless. but still we won this game #worldcup
```

Figure 4 shows the first semifinal, which includes a number of goals, yellow cards, references to Paul the Psychic Octopus (a popular character in the news at the time with perfect record on predictions) and the final keyword "final" that summarizes the fan reports that one team has reached the WC finals. Note the presence of players' names from both teams as they were fundamental in the development of the game. Table 2 presents a few keywords and corresponding tweets.

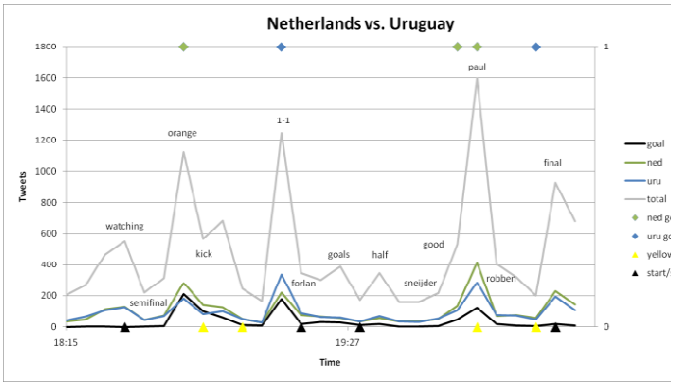

Figure 4. Game #61: Netherlands vs. Uruguay. A back and forth game with goals and high activity from both teams.

Keyword	Tweet
Orange	ORANGE SCORE! wow what a goal! 1-0 #NED #worldcup
	Goal! #ned #worldcup "the future may be orange"
Forlan	Correction: 24' is 24 feet. Forlan's shot was from 24 yards out, or 72'. So there. #WorldCup
	Forlan. Blimey, Alex Ferguson must be thinking he's a different player to the one he signed so long ago. Such power in that boot. #worldcup
Sneijder	Goaaaaaaaaaaaaaal...scrappy but much needed and welcomed....yipee #ned #worldcup... #uru is goin down. Goal no 5 for Wesley Sneijder
	Goal!! Sneijder is the man!! Hup Holland Hup! #ned #worldcup
Paul	lol paul the octopus chose #esp to win the #esp #ger match. #worldcup that damn octopus has been 5 for 5
	I don't trust octopus Paul anymore… he has no idea how were gonna beat Spain 2morrow! :D #GER will win the #worldcup :)

Table 2: Examples of keywords and tweets for Figure 4.

The second semifinal was less interesting in terms of action (i.e., fewer goals, no cards, etc.). One of the most exciting moments was at the very start of play, when a fan ran onto the field with a vuvuzela and had to be escorted off by security. Around middle of the second half, "boring" became a popular keyword. Once the game ended, the discussion promptly moved on to anticipation of the final, with "Holland" (Netherlands) becoming a spiking term. Table 3 presents a few keywords and corresponding tweets. The game for the 3rd place is presented in Figure 6. Note again the pattern of players' names on keywords. This is expected as the crowd, that has been following games for near a month, is now very familiar with the top players. The crowd still has a fresh memory of the `#URU` vs. `#GHA` game and expressed displeasure, shown as "booing", to a player even though he is playing well.

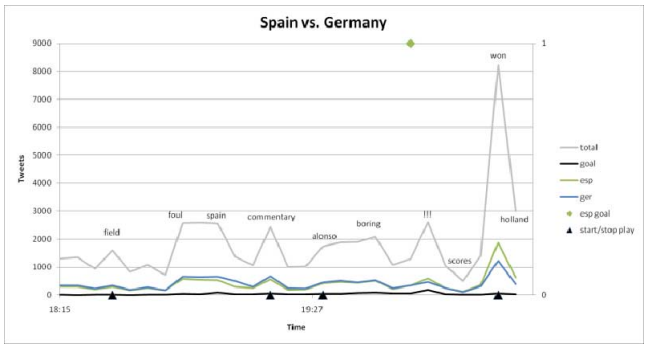

Figure 5. Game #62: Spain vs. Germany. Only one goal in a match that apparently failed to meet expectations

Keyword	Tweet
Foul	#WorldCup – LOVE watching Spain and Germany play… 20 mins in, not a single foul. Beautiful, clean soccer! #ESP #GER
	Has there been any sort of foul already in this match? Ah… someone offside… #gerspa #worldcup
Boring	For a semi-final, bit of a boring match,unlke the ova semi-final game where #Holland OWNED!Lool !O DID U C THAT FAIL? come on #ESP #worldcup
!!!	YES!!! #worldcup #esp
	Germany wake up!!! Stop watching and start fighting!!! #worldcup

Table 3. Examples of keywords and tweets for Figure 5.

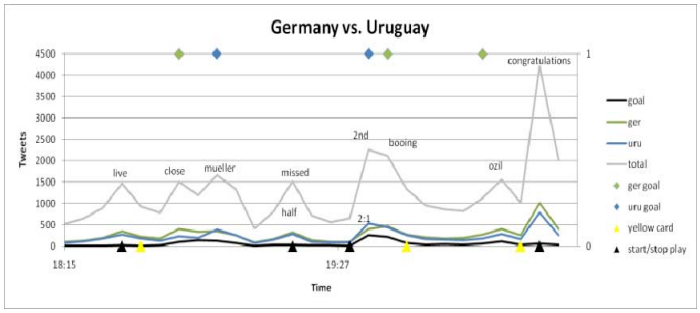

Figure 6. Game #63: Germany vs. Uruguay. Vibrant game with goals and lots of activity.

Keyword	Tweet
Missed	One half in at Nelson Mandela Bay Stadium. It's all tied but both teams have missed chances to open it up. #URU 1 – 1 #GER #WorldCup
	DAMN German dude just missed a one timer that was sure to go in before half- 1-1 in a great game. #worldcup
Booing	GOAL #URU. Forlan is incredible player. #worldcup. Crowd still booing Suarez.
	Despite rooting for Uruguay, Suarez deserves all the booing. #worldcup
Congratulations	Good #WorldCup match, congratulations to Germany for obtaining the Bronze Cup.
	Oh.unlucky Diego. And congratulations #ger for reaching the "podium" for the third #worldcup in a row. Final: Uruguay 2, Germany 3.

Table 4. Examples of keywords and tweets for Figure 6.

5. EVALUATION

We performed two types of evaluation. In the first one, we compare against a world-famous sports section that reported extensively about the event. The second examines the quality of the timelines using crowdsourcing.

5.1 Expert Editors

To evaluate the timelines, specifically our keyword detection, we compared against BBC[2] live blogging of the same events. We selected BBC because of their very detailed coverage of the event, expertise in the domain, and reputation. We believe the BBC Sports editors represent a much stronger baseline than untrained human editors at identifying noteworthy events in the game.

Some excerpts from a live blog are given in the table below. Since the full text of the BBC live blog was much shorter than the text produced on Twitter, but presumably more definitive, we computed the % of terms in the BBC coverage, not including stop words, that were also present as spiking terms on Twitter at around the same time as the post, where spiking terms are defined as any with a z-score of at least one during the same time interval. Thus, for each 5min. interval in which the BBC journalist posted commentary, we produced a score between 0 and 1. By summing the counts of overlapping and editorial words over the game, we arrived at an overall score for each game as well.

Timestamp	Comment
14:00	KICK-OFF Germany v England Germany and England's last-16 World Cup clash is under way in Bloemfontein. Here we go.
14:20	GOAL Germany 1-0 England That's among the worst England defending I've seen at a World Cup, I'm gobsmacked. Manuel Neuer's goalkick sails over John Terry's head and Miroslav Klose shows great strength and coolness to hold off Matthew Upson and poke past David James. Klose could have gone down there and Upson would have been sent off, but he bagged his 50th international goal instead.
15:14	Our friends at Infostrada Sports tell us that Germany have played 30 previous World Cup matches in which they took a 2-0 lead. They have won 29 of them and lost one: in the last 16 round in 1938 against Switzerland (2-4). That is 72 years ago.

Table 5: Excerpts from BBC live blog. Words also spiking on Twitter are highlighted in green; stop words are gray.

For the Germany vs. England game, for example, our evaluation indicated of 14.8% coverage of editorially produced terms in the Twitter traffic. While this seems like a low number, it includes all non-stop words of the BBC coverage. Function words, in particular, do not tend to appear as spiking terms, and the journalistic coverage is sometimes only tangentially relevant to the game, as shown in the third example post in Table 5. To simulate non-timely coverage of the same material, we offset the times in the Twitter traffic by an hour to simulate irrelevant coverage, again using the Germany vs. England game's Twitter traffic and BBC coverage. With this data, the coverage score dropped to 5.9%. We also simulated irrelevant commentary by evaluating the Twitter traffic for the Germany vs. England game against the live blog for the Slovenia vs. United States game. The coverage score in this case was 5.2%, with the overlapping terms consisting of generic soccer terms such as "ball" and "keeper" and some common words missed by the stop word filter, including "just" and "want." These coverage scores, as well as the scores for all 6 games presented here, are summarized in table 6.

Game	Coverage	BBC Words	Twitter Words
Germany vs. England	14.8%	608	1,501
Slovenia vs. United States	8.6%	498	1,526
Slovakia vs. Italy	11.3%	433	620
Netherlands vs. Uruguay	5.9%	545	183
Germany vs. Spain	13.3%	534	917
Germany vs. Uruguay	12.4%	539	648
Not timely	5.9%	608	1,211
Irrelevant	5.2%	498	1,403

Table 6. Coverage scores for World Cup games and baselines

The quality of our summary, as measured by this evaluation, varied over the course of the game. This is reflected in figure 7, which shows the coverage score, the total tweet volume, and the times of events. Total volume and event times are analogous to those shown in the graphs in section 5.5. As the graph shows, coverage is often best during periods of lower Twitter volume or around the events that we have explicitly noted on the graph. At the objective events, the overlap likely stems from keywords describing the event, e.g. "goal", or aspects of the event, such as the resulting score or the names of the players involved. Why the coverage improves during periods of low volume is less clear. This might mean that the spikes, at least to some degree, reflect the flow of the conversation on Twitter rather than strictly the rhythm of the game.

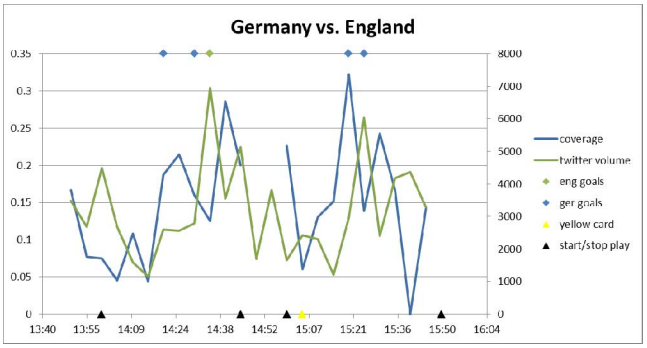

Figure 7. Evaluation of Game #51: Germany vs. England.

5.2 User Study

We conducted a study using crowdsourcing via Mechanical Turk. The goal of the experiment was to assess the overall quality of the timeline summaries. The experiment was self-contained with minimal instructions and no qualification test was conducted prior to judgment. Each HIT was assessed by 7 workers and the pay was $0.10 per assignment (total cost $4.26). The task consisted of presenting the users with an image of the timeline summary (the exact same figures 1-6) and the following questions:

[2] http://news.bbc.co.uk/sport1/hi/football/world_cup_2010/fixtures_and_results

1. What was the final score of the game?
2. Based on the summary, please rate the game on a scale 1 to 5 (1=boring, 5=exciting)
3. Is there anything in particular that you found interesting in the timeline summary?
4. Please rate how informative the timeline summary is on a scale 1 to 5 (1=irrelevant, 5=outstanding)

Game	q1 (percentage of correct answers	q2	q4
#61	71%	4.14	3.14
#62	86%	2.86	3.43
#22	100%	4.43	3.86
#51	86%	4.14	3.71
#63	100%	3.71	4
#41	83%	3.5	3.67

Table 8. Results from the experiment using crowdsourcing.

The results are summarized in the Table 8. The games were presented in random order. Most the workers were able to report the correct score by looking at the picture (column q1). However, there is an indication that the choice of colors may not be the most appropriate ones. Columns q2 and q4 show the average results for questions 2 and 4 respectively. Note that for game #62, workers clearly realize that the game was not very interesting. Below is a sample of un-edited comments for question 3. One can see the workers were able to pick interesting bits by visual inspection.

Game #61: *it s awesome to compare the performance*

THE TWEET ACTIVITY COUNT REGARDING THE PREDICTIONS OF PAUL WAS OUTSTANDING IN THE SECOND HALF

Game #62: *Looks like the game was fairly dull with little incident which may explain why there was such a peak of tweets at the full time.*

Low interest in general, and compared to sample timeline no interest that caused spikes of tweets other than "won" at the end where little spikes during game

Game #22: *Americans getting upset and tweeting "crap" after Slovenia's first goal.*

It's funny that the tweets say "finally" when US got the goal

Game #51: *I thought the wwii tweets were perplexing since England at least was able to win wwii*

How the response to Germany taking a 2-0 lead was so much greater than the response to England's wrongly disallowed goal.

Game #63: *It explained tweet activity, which I have never seen before and is very interesting.*

Diego Forlan wasn't in it even though he had a good goal

second half was more exciting

Game #41: *People seemed happier about Italy getting eliminated than Slovakia winning.*

short and sweet way of telling whole thing

The performed evaluation shows that our design has potential.

6. GENERALIZATION

In addition to our detailed study of the World Cup, we produced timelines for a number of other datasets. In particular, we can easily apply the same method used for soccer to other sports. First, we analyzed Superbowl 45, from February of 2011 (figure 8). The major features of the game are analogous to the soccer games, though because the game is longer the dip in attention during the second half is more noticeable. The detected keywords include score reports and the names of players, who contributed to scoring, were injured, or participated in a bad play. Limitations in applying our visualization to this dataset include difficulty in representing multiple score changes in one time interval, as happens in American football with conversions after touchdowns. In addition, as opposed to soccer where each goal is one point, changes in score in football can vary in magnitude. We could perhaps add a second dimension to the score line in our visualization, but the differences in magnitude are not representable within our current interface. An additional challenge with this dataset is the large volume of traffic commenting on the commercials, rather than the game. The keyword "updates," for instance, is a comment on a Chevy commercial for a car that displays Facebook updates from your friends on the dashboard. In many time intervals, especially ones in which nothing major occurs in the game, keywords from the commercials dominate.

The second data set was the 2011 Rugby World Cup (`#rwc2011`). The timeline for the final match of the rugby world cup (New Zealand vs. France) is presented in Figure 9.

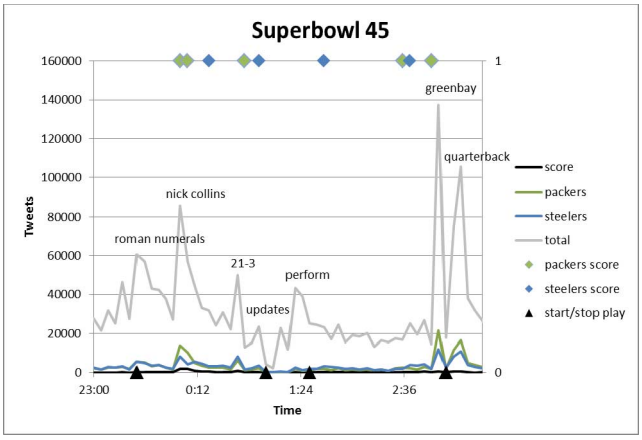

Figure 8. Timeline for Superbowl 45

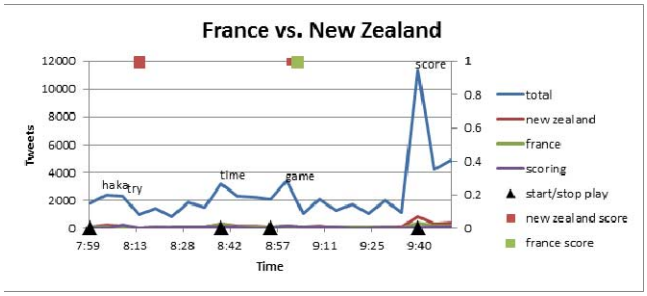

Figure 9. Final game between New Zealand and France.

7. CONCLUSIONS AND FUTURE WORK

Without doubt Twitter has become a very important source of information about world events. In the case of popular scheduled

events like a sports game, the level of engagement is quite high and shows that the public are interested in following what's going on and adding their own comments to the discussion. However, as presented by others, it is sometimes is difficult to make sense of all the information coming through the fire hose.

In this paper, we presented a timeline generation design that allows us to capture the most salient events of a game along with the most popular keywords as annotations alongside a timeline. The motivation behind the design is to reflect a bar metaphor where subjective comments are mixed with the outcome of the game. For sports fans, this type of information is very relevant as they are interested specific details and the overall conversation.

Based on our findings, the crowd follows the game with an obvious human delay effect but, in general, is able to produce an accurate description of the outcome of the game. It is then possible, given the right data mining tools, to follow a game completely on Twitter without missing any of the important facts. In fact, microblogging may give you other information that traditional media are not reporting. Our evaluations show that our technique is sound and that users are able to understand the basics. However, there is also a potential indicator that such timeline summary may be more appropriate for more informed fans and therefore more studies need to be completed to make the timelines more informative to casual observers of the games.

Ongoing work on this topic includes improving the presentation and visualization of the display based on the feedback from the user survey. Geotagging information was not used in this project and is something we plan to incorporate for upcoming events. Algorithmically, we would like to extend our keyword detection to use higher-order n-grams and produce summaries in real time rather than from static, historical datasets. At present all of our algorithms are already online algorithms except for keyword detection, which could easily be adapted by simply excluding future tweets from the calculations of mean and standard deviation. The keyword detection might also benefit from using time-based weights in the mean and standard deviation calculations to make the scores of keywords depend more on temporally close tweets. A more concrete evaluation, perhaps formalizing our event detection so we can predict where goals occurred and compare against the actual times of goals, could help us tune parameters as our keyword detection becomes more complex.

Finally, we are interested in deploying these timelines in an exploratory search system where the users can interact more with the conversations, perhaps by integrating the example tweets and some form of drilldown into the timeline visualization. Part of any changes to the interface would likely involve solving the problems presented by generalization to domains besides soccer, such as choosing which and how many volume lines to show and, for sports with more frequent or varying magnitudes of scores, finding a non-binary way of representing changes in score.

8. REFERENCES

[1] J. Allan *et al.*Temporal Summaries of New Topics. In *Proc. SIGIR*, 2001.

[2] O. Alonso *et al*. Time-based Exploration of News Archives. In *Proc. of 4th HCIR*, 2010.

[3] H. Becker *et al.* Beyond Trending Topics: Real-World Event Identification on Twitter, *AAAI* 2011.

[4] D. Chakrabarti and K. Pundera. Event Summarization using Tweets, *AAAI* 2011.

[5] M. Dork *et al*. A Visual Backchannel for Large-Scale Events, *IEEE Transactions on Visualization and Computer Graphic*s, Volume 16, Issue, 6, (2010).

[6] D. Gaffney. #iranElection: Quantifying Online Activism, *Web Science Conference*, 2010.

[7] J. Hannon *et al*. Personalized and Automatic Social Summarization of Events in Video, *IUI* 2011.

[8] J. Huang and M. Iwaihara. Realtime Social Sensing of Support Rate for Microblogging, *DASFAA Workshop*, 2011.

[9] A. Kokaram *et al*. Browsing Sports Video, *IEEE Signal Processing Magazine*, (47), March (2006).

[10] J. Makkonen and H. Ahonen-Myka. Utilizing Temporal Information in Topic Detection and Tracking. In *Proc. ECDL*, 2003.

[11] J. Makkonen *et al.* Topic Detection and Tracking with Spatio-temporal Evidence. In *Proc. ECIR*, 2003.

[12] A. Marcus *et al*. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI* 2011,

[13] M. Matthews *et al*. Searching Through Time in the New York Times. In *Proc. 4th HCIR '10*, pages 41-44, 2010.

[14] A. Popescu and M. Pennacchiott. Dancing with the Stars, NBA, Games, Politics: An Exploration of Twitter Users' Response to Events. In *Proc. ICWSM,* 2011.

[15] A. Qamra *et al.* Mining Blog Stories Using Community-based and Temporal Clustering. In *Proc. CIKM,* 2006.

[16] B. Shaparenko *et al*. Identifying Temporal Patterns and Key Players in Document Collections. In Proc. *IEEE ICDM TDM Workshop*, 2005.

[17] D. Shamma *et al.* Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events?, *CSCW* 2010.

[18] R. Swan and J. Allan. TimeMine: Visualizing Automatically Constructed Timelines. In *Proc. SIGIR*, 2000.

[19] H. Takamura *et al.*, Summarizing a Document Stream, In *Proc. ECIR,* 2011.

[20] D. Tjondronegoro *et al*. Multi-Modal Summarization of Key Events and Top Players in Sports Tournament Videos, *IEEE WACV* 2010.

[21] S. Wakamiya *et al.*, Towards Better TV Viewing Rates: Exploiting Crowd's Media Life Logs over Twitter for TV Rating. *ICUIMC'*11.

[22] C. A. Yeung and A. Jatowt, Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining, In *Proc. CIKM* 2011.

[23] S. Zhao *et al*. Humans as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. Technical Report TR0620-2011, Rice University and Motorola Labs, June 2011.