Crowdsourcing for Football Match Summaries

B Sai Kiriti[1] and Aritra Ghosh[2]

[1] Indian Institute of Technology, Madras
kiritib@cse.iitm.ac.in
[2] Indian Institute of Technology, Madras
aghosh@cse.iitm.ac.in

Abstract. With the explosion of the amount of data being generated by Micro-blogging services like Twitter every day,there has been a wide spread and increased interest in making sense of unstructured data which in the case of Twitter is highly repetitive and noisy. In this paper, we study the problem of summarization of planned events in specific football matches. We propose a technique which generates highlights of a given football match very similar to the summary. Temporal spikes in the volume of tweets are used to identify important moments within an event and Ranking words based on Top-k discriminating words Ranking followed by diversification using Jaccard Similarity approach has been used to extract relevant and representative tweets from the important moment within an event. Evaluation of the proposed methods on tweets obtained from 2-3 football matches of the 2010 World Cup generate summaries which are readable, factual and competitive with human tailored summaries of the same events.

Keywords: Social Media, Text Summarization, Microblog

1 INTRODUCTION

Twitter has turned out to be the most popular micro-blog website with more than 0.5 billion and 0.284 million active users[1]. Every day on Twitter, humongous number of tweets(200 million updates according to [2]) are being posted. Many of the tweets are about events that people describe and concern with natural disasters[3],regional riots and scheduled events like political debates, local festivals and sporting events [4]. Twitter streams thus cater to a varied diverse and broad range of events and broadcast this information in real time. However there are certain issues for these to be accessible to humans. Given the large rate of tweets every moment, it is difficult for the user to follow the full stream. Secondly, a lot of the tweets consist of spam and redundant in their content. Thus there is a need of automated summaries which would be consumable to an user without a person manually generating a summary or highlight or certain news stories which journalists are unable to cover.

In this work , we design an automated system to crowd-source summaries of events using tweets posted as the only source. In specific , we focus on summarizing football matches of World Cup primarily because they are scheduled events and are associated with large volumes of tweets due to world wide craze and also there is good amount of press coverage to serve as gold standard. Our method works in a 2 stage manner- In the first stage we identify the important moments using a frequency based approach where we identify the peaks in the tweets frequency graph using several intuitive and efficient heuristics. Once the important moments are identified, we need to extract the representative tweets corresponding to each important moment. This is done using a Ranking score for each tweet within a moment based on certain discriminating words in the tweet. We would additionally also like the representative tweets within each moment to be as diverse as possible.Interestingly, our method works in an unsupervised way with very minimal manual keywords about the event encoded. This can thus be extended very easily to other scheduled events without encoding additional knowledge.

2 RELATED WORK

Although a lot of work has been done in text summarization domain, these techniques cannot be extended to microblog summarization due to reasons discussed above(noisy and repetitive nature). There has been a renewed interest in the field of Microblog summarization especially after the work done by Sharifi et al[5][6]. Chakrabarti et al [7] make use of Twitter to generate summaries of long running, structure rich events under the hypothesis that multiple event instances share the the same underlying structure. They model the structure and vocabulary of events of American football using a Hidden Markov Model(HMM). Some of the issues here are the time taken by the HMM's to learn and the supervision required(Tweets of many previous matches).
Nichols et al([8]) detect important events and a journalistic summary from tweets from a World Cup football game. They use temporal cues to identify the important moments within an event and within an event and then rank the sentences by constructing phrase graph based on scores to nodes.Louis and Newman[9] proposed a method for summarizing a collection of tweets related to a business. The proposed procedure aggregates tweets into subtopic clusters which are ranked and summarized by a few representative tweets from each cluster. Some other methods require domain knowledge or manual supervision making them infeasible.

3 METHODOLOGY

3.1 Tweets for Football Games

Collection of Tweets The dataset we use for the evaluation of our method is from Nichols et al[8].The dataset was collected from tweets of 36 games of the 2010 World Cup. Twitter's Streaming API was used to extract tweets based on keyword query. They were recorded using words like "worldcup" and "wc2010" which were promoted by FIFA(www.fifa.com) and Twitter for World Cup games.

Since some of the games occurred simultaneously, keywords such as the name of the countries , country abbreviations, and nicknames of the countries were used in conjunction . Much of the evaluation will require annotation from humans and thus we work with 3 of the 36 games. Among the 3 we focus most of the results on the US vs Slovenia on 18th June, 2010. However ,we note that the method works for any of the games with minimal amount of manual encoding. Some of the statistics appear below .

Table 1. Statistics of the Football Dataset

Game	Total number of tweets	Mean Tweets/min	Min Tweets/min	Max Tweets/min	σ
US vs Slovenia	113189	752.26	300	3992	572.25
Germany vs Serbia	72335	478.5	151	1810	309.7
Australia vs Serbia	224046	1427.04	315	4459	840.38

Nature of Tweets Figure 1 shows the tweet volume graph which recorded the number of tweets per minute for the game between US and Slovenia . We note that time span of the graph is more than the actual duration of the match. This

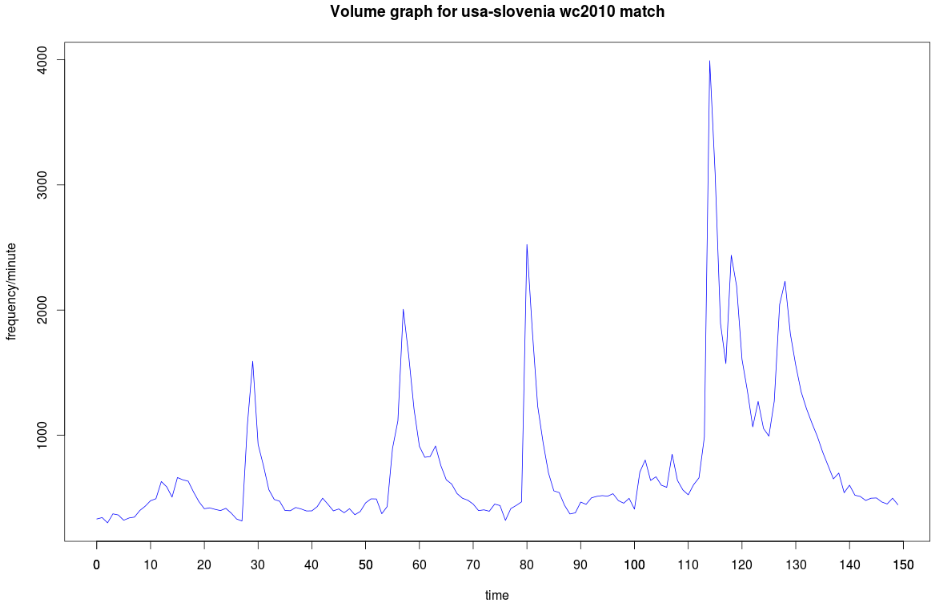

Fig. 1. Tweet Volume Graph for the US vs Slovenia Match

is because the dataset contain some tweets(for a small duration) both before the match and after the match. After examining some of the tweet volume graphs we find some of the common features among them which are listed below

- There are several spikes in the graph and most of them correspond to various important moments in the match like goals, penalties, red card etc.
- There is a lag between the actual event and the spike in the tweet volume graph. This follows from natural intuition since people tweet only after a certain time.
- The contents of the tweets contained in a spike can be very diverse as well as repetitive and can have multiple sub-events Eg: Penalty followed by Goal or Goal followed by red card.

3.2 Summarization System

A football match consists of a ordered sequence of moments some more important than other , each of which contains actions by players, the referees, etc. The input to the algorithm would be a stream of tweets of a Football match and the output would be summary i.e in some sense the highlights of the game. We note the summary would be a subsets of the original tweets. Thus we do not expect a journalistic summary but summary on similar lines which gives us maximum information. On a high level the following are the steps in order to extract the summary of the match

- From the Tweet Volume graph , various important moments are detected using various heuristics based on slope of the graph.
- Noise Elimination in terms of removal of spam , non-English etc is done.
- Representative diverse tweets from each moment are computed using a top-k words Tweet Ranking based approach and diversification based on Jaccard Similarity.

Detection of Important events For the detection of important moments we adopt an approach similar to [8]. However, we make several modifications. The tweet volume is in the granularity of minute. The algorithm we use is an offline algorithm i.e we use statistics from the entire match. We choose peaks as those points where the slope of the Tweet Volume graph changes from positive to negative and the absolute number of tweets posted exceeds a certain threshold. The threshold is an empirical parameter and we have tried several approaches for the same. One of the naive ways would be visually inspect the graph and set it. Another intuitive is to consider only those peaks which have number of tweets greater than mean(μ) + 2*standard deviation(σ) . However one issue with this approach is that mean is heavily skewed by outliers and thus proves to be detrimental. On this line , the authors in [8] use a threshold of 3*median . However we see that in our case , it does not work very well. We use median + c*standard deviation(σ) where we choose empirically. From our experiments we see that $c = 0.5$ works very well. It would be interesting to choose the threshold in a more systematic way. We also tried finding the threshold using a window approach ignoring the peaks.However it failed to capitalize.

Selection of Tweets from Important Moments Once we have detected the important moments, we need to decide the tweets we need to consider for each of the important moments from which we need to generate the summary. We note that a very large spread of tweets will lead to lot of false positives(tweets that do not give important information) and a narrow spread of tweets will lead to true negatives(missing out on important tweets). We choose a k minute interval around the peak i.e all tweets in the interval $[T_p - k, T_p + k]$ where we empirically see that $k = 2$ gives us the best results. In [8], the authors try an approach to extract those tweets within the spike that have a high Signal to Noise Ratio(SNR).It is important to note however that the interval need not be symmetric about T_p. We also tried an approach of choosing all tweets from where the slope begins to rise (based on some threshold) to where slope falls almost to 0. This approach works well although we use the former.

Noise Elimination This step is very important in the process of generating summaries. We first remove all tweets which have less than 3 correct English words since even if the language specified for that tweet by Twitter is English, the tweet may not actually be in English. This also takes care of short tweets that are not likely to give us any additional information. Spam is removed using a dictionary of common terms that are found in tweets and the stop words from nltk package .Some other ways include removing all tweets which have urls in them, have pronouns like I,me,Our. Here we base ourselves on a strong assumption that the signal from tweets dominates the noise which is generally the case.

Choosing Representative tweets Here, we first find the top-k words(based on frequency) that appear in the specific moment. We also store the corresponding frequency.From the top-k words list(T) we remove words of length 1 and 2 and words which do not have discriminating power like goal, the names of the two teams etc. Now ,we rank the tweets based on the score of each word in the tweet where

$S(t) = \sum_{i=1}^{n} score(t_i)$ where $score(t_i) = freq(t_i)$ $t_i \in T$

We note that score of a word is 0 if it does not belong to the top-k words list. The top-k words chosen in this method do not comprise of spam words and other common words like goal, team name etc. We take top k' tweets(S_{rank}) based on $S(t)$. Empirically we see that $k = 20$ and $k' = 20$ works best. This is to ensure that we take only those tweets which give us some additional information and serve as representative of tweets in the moment. In addition, this helps in avoiding spam teams which are off topic. However on obtaining this list of tweets we see that many of them are very similar and have the same content. Therefore we need to choose diverse tweets from this. For this we use the notion of Jaccard Similarity of tweets .

$$J(t_i, t_j) = \frac{|S_i \cap S_j|}{|S_i \cap S_j|}$$

i.e the total number of common words to the total number of words in both the tweets.Here S_i refers to the set of words in the tweet t_i Now we take all sets of k tweets from given n tweets(S_{rank}) from the Ranking method and choose the set that has the minimum similarity Sum where Similarity sum is defined as .

$$SS(K) = \sum_{t_i, t_j \in K and i \neq j} J(t_i, t_j)$$

.

This ensures the that the set of tweets are diverse and do not give redundant summaries. Again here we choose $k = 3$ which is a good choice in our case i.e 3 tweets summarizing every important moment. The number of tweets summarizing each important moment is dependent on the specific application and can be changed accordingly

4 EXPERIMENTS AND RESULTS

In this discussion we describe the various evaluation metrics used and the results obtained.

4.1 Evaluation of Important Moment Detection

To evaluate the performance of moment detection we compare the counts of various types of key moments given by our algorithm to those obtained by reading from recap articles. Like in [8] The following categories are chosen for key events : goals, penalties, red cards, yellow cards,disallowed goals.

The event counts for the algorithm were gotten by manually examining all the tweets which were chosen from each important moment. Here we should note that a moment may contain more than one important sub-events as discussed earlier.

The actual event counts are obtained from www.guardian.com, en.wikipedia.org, www.goal.com.Using more than one source ensures that none of the key events are left out.We also use some of the sources as the Gold standard.

We use the standard definition of Recall(R) and precision(P). Recall is the ratio of number of key events detected by the algorithm to total number of events which actually occurred. Similarly, Precision is defined as the number of key events detected by the algorithm divided by the total number of events returned by the algorithm. They are summarized in the below table We see that recall is less as compared to the precision. This is because a key moment returned by the algorithm will most likely be a important event. We see that in general yellow cards, red cards and disallowed goals are difficult to detect since they are unlikely to create a spike in the tweet volume graph. The Germany vs Serbia

Table 2. Recall/Precision Measure of Moment detection

Game	Actual Moments	Moments detected	R	P	F-measure
US vs Slovenia	10	7	0.7	1	0.824
Germany vs Serbia	11	8	0.54	0.875	0.67
Australia vs Serbia	8	11	0.625	0.45	0.52

match had 9 yellow cards in actual. As discussed , some of the yellow cards were not detected by our algorithm leading to a fall in the recall. We see that in case of the Australia vs Serbia Match many of the tweets contain tweets relating to Germany vs Ghana match which is because both were held simultaneously. This reduces the precision drastically.

4.2 Evaluation of Summarization

Most of the analysis here is subjective. Although ROUGE-N metric is used in general text summarization problems ,we do not use it here because of the lack of a "true" gold standard . Nevertheless, we do compare it with a manual summary. The comparisons are shown in the table. In general the recap articles do not contain a moment wise summary but only a overview of the match. Therefore we are forced to resort to manual summary. The manual summary used here is same as in [8]. We see that most of the times the recap articles are longer

Table 3. Comparison of Summaries generated vs Manual Summary

Game	Spike Index	Manual Summary	Generated Summary
US vs Slovenia	1	In the first 15 mins of the soccer game between USA and Slovenia, Slovenia is leading with a goal by Birsa. Birsa scored an easy goal from midfield to the right of the goal, as USA left that shot wide open. Terrible defense by USA team, too much space left open	and birsa gives #svn the lead. what a goal! 20+ yards into the top corner. #svn 1-0 #usa #worldcup', "what happened tim howard? you didn't even move. ugh. good job slovania on the good goal. but come on #usa!! #worldcup", 'slovenia scores, a great game so far, seems like a perfect shot from birsa //john #usa #snv #worldcup'
Germany vs Serbia	1	Klose argues with referee, gets second yellow cards and is out of the game. Germany down to 10 men.1-0 Serbia	'red card in #ger game wow klose got a red!!!!!!! #worldcup #ger down to 10 men vs #srb', 'this referee is a joke. germany down to 10 men. klose sent off. and serbia take the lead. #ger 0-1 #ser #worldcup', 'and milan jovanovic scores for serbia! 1-0 with seven minutes until half time, and this is disastrous for australia. #worldcup #srb #ger'

than the summaries we obtain by our algorithm. In general, we see that the summaries we get are meaningful and have a good coverage as well as factual and not redundant in terms of the content generated.

Statistics of Match In addition to the above, we also tried generating several statistics of the match. They are summarized below.

- Score of the Match: To find the score of the match we search for the regular expression "$\backslash d - \backslash d$", "$\backslash d : \backslash d$" in the last fifteen minutes of the match. Although the same could be found over the tweets concerning the whole match, we avoid the route to avoid additional computations.
- Timings of Yellow Card : This is found by plotting the Frequency of tweets with the keyword "yellow" or "#yellow" or "yellow card" and noting the peaks in the graph.
- Red Cards: Same as Yellow Card with appropriate replacements
- Penalty : Similar to the above we search for the keyword "penalty" or "#penalty" and plot the Tweet volume graph.

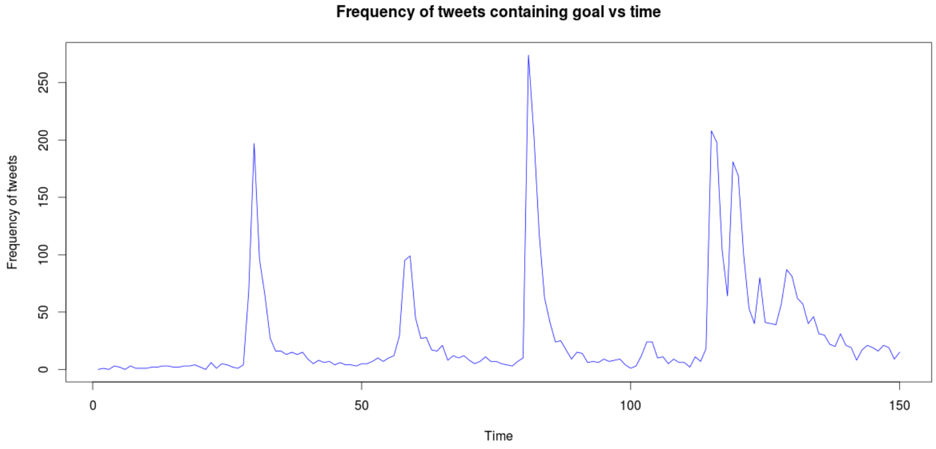

Fig. 2. Tweet Frequency Graph for tweets containing keyword "goal" for USA vs Slovenia Match

In Figure2, We observe that there are broadly 3 clear peaks and one is spread out over time. In reality, the three peaks corresponds to goals and the last one is also a goal. However the spread out in the peak is because just after there was a goal which was disallowed leading to many tweets posted about it . We see that the above methods are very effective and we have almost always 100% recall as well as precision in detecting the above statistics. To conclude ,we also find the players who scored the goal by finding the frequency of all the words occurring around that moment(when the goal was scored). We expect to see that the frequency of the person who scored the goal to be occurring maximum number of times(considering only nouns).

5 CONCLUSIONS AND FUTURE WORK

In this paper, We have proposed an algorithm for the summarization of Football Match using crowd sourced tweets . Our algorithm mainly is based on the premise that a spike in the volume of tweets posted corresponds to a important moment which is generally true. Then using a top-K words approach, we rank the tweets and choose a subset such that they are dissimilar among themselves. We obtain reasonable summaries which are intelligible and factual and on par with human generated summaries.

One aspect we have not looked as a part of our work is how augmenting information from previous matches would help better the summaries and also the detection of important moments i.e we would like to explore supervised approaches for solving the problem. Another issue we envision is that the method has a number of empirical parameters and an issue may arise when the domain is changed. We would like to see how the method works for other sports. Lastly, the evaluation methods used are adhoc and do not allow for comparison between different methods neither give us a good idea of the effectiveness of the method proposed. We would like to study how better evaluation measures can be formulated for the problem .

References

1. http://www.cs.technion.ac.il/ gabr/resources/data/ne_datasets.html
2. https://blog.twitter.com/2011/200-million-tweets-day
3. Kate Starbird, Leysia Palen, Amanda L Hughes, Sarah Vieweg: Chatter on the red: what hazards threat reveals about the social life of microblogged information Proceedings of the 2010 ACM conference on Computer supported cooperative work Pg 241-250(2010)
4. Hannon, J., McCarthy, K., Lynch, J., Smyth, B. Personalized and Automatic Social Summarization of Events in Video,*In Proc IUI 2011*
5. Sharifi,B., Hunton, M.A., Kalita,J. Summarizing Microblogs automatically Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics Pages 685-688
6. Sharifi,B., Hunton, M.A., Kalita,J. Experiments in Microblog Summarization Social Computing (SocialCom), 2010 IEEE Second International Conference Pg 49-56 .
7. Deepayan Chakrabarti , Kunal Punera Event Summarization Using Tweets. ICWSM , 2011
8. J Nichols, J Mahmud, C Drews Summarizing sporting events using twitter. Proceedings of the 2012 ACM international conference on Intelligent User
9. Annie Louis and Todd Newman. 2012. Summarization of business-related tweets: A concept-based approach. In Proceedings of the 24th International Conference on Computational Linguistics (COLING)