Twitter Analysis of IPL cricket match using GICA method

Ajay Ramaseshan, Joao Pereira, Santosh Tirunagari

July 28, 2012

Abstract

Twitter is a powerful medium to express views and opinions, in fields such as entertainment, sports and politics. In this project a twitter analysis of a cricket match was done with the help of GICA method to analyse the tweeting patterns of users. The tweeting patterns were then correlated with the match statistics to understand the similarities and contextual information in the tweets.

1 Introduction

GICA method [1] is a powerful tool to study contextual information and pragmatics of language. Pragmatics is the study of the context of the speaker. For instance, let us take a simple phrase such as "green light". The implicit meaning or lexical meaning is a light which is green in colour. However, in this sentence "green light to the project", the context of "green light" is a go-ahead. This understanding of context is what GICA method aims to do. Generally, to analyse the statistical occurence of keywords in a document, the term-document frequency is calculated and a 2 dimensional matrix is obtained. The GICA method extends this further into a 3 dimensional tensor with Contexts x Subjects x Objects. The subjects and objects correspond to the documents and words respectively, while contexts is a higher order understanding of the text. The context could be related to environment, time, main theme of the text and so on. For example, if words such as benzene, toluene, aldehyde and alkene occur in a document, then the context would be Organic Chemistry. In this work, the GICA method has been applied to the twitter analysis of IPL tweets.

The Indian Premier League or IPL [2] is an annual cricketing competition that is held in India from April-May every year. Eight to ten teams play each other in a round robin league stage, after which the semifinals and finals are conducted. It is a widely watched event in India and a lot of enthusiasm surrounds this contest. The official twitter channel of the IPL [3], sees a lot of tweets during this season. These tweets reflect the opinion of the public about the performance of the teams to a large extent. Thus we have a huge collection of contextual and pragmatic information in these tweets and thus the GICA method could be applied to analyse the contexts within these tweets.

2 Methodology

To analyze twitter feeds we gathered corpus by collecting tweets from different Twitter users over the duration of a single IPL cricket match. Each tweet was

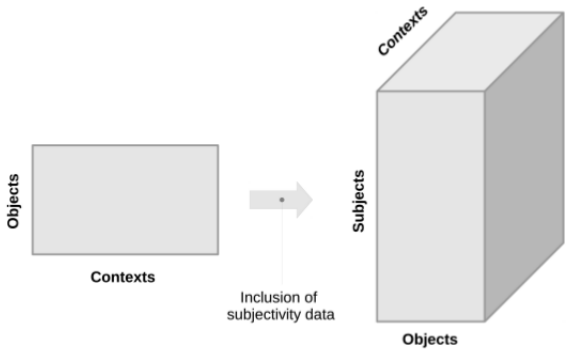


Fig. 1. An illustration of an object-context matrix expanded into a tensor that accounts additionally for subjectivity. In other words, we perform an extension of a $O \times C$-element matrix into a third-order tensor of $S \times O \times C$ elements, where the data concerning different subjects on objects in contexts are included.

achieve	ball	bat	bowl	captain	catch
cricket	fan	game	ground	hatrick	league
match	over	patnership	score	wicket	win

Table 1: Context words.

handled as a separate entity. We analyse the similarities and differences between the tweets and try to correlate the information gathered from the twitter analysis to the match statistics. The GICA method is now applied over the collected corpus. The data is converted into a 3 dimensional tensor

$$X \in R^{CXOXS} \tag{1}$$

of Contexts x Subjects x Objects [1]. Cricket being the main context or the environment, thus the contexts of interest would be the different cricketing terminologies such as ball, bat, over, runs and so on. The subjects are the two participating teams of the match. The objects are obtained by collecting the top 25 most used words in the whole corpus, excluding stopwords like articles and expressions in everyday informal language such as "imba,lol" which are quite common in tweeting [1]. This is illustrated clearly in Figure 1 [1] shown above.

In order to analyse this tensor, the 3 way tensor can be flattened into 2 dimensional matrices. As discussed in [4], the fibre of a tensor is the subarray consisting of elements having all indices except one fixed. Thus, the fibres of the Contexts would be the O X S (Objects x Subjects). Matricization could be done in each of the three dimensions. In the first case we get a 2 dimensional matrix with Contexts as rows and Objects X Subjects as columns. Matricization along the Objects produces a 2 dimensional matrix with Objects as rows and (Context X Subjects) as columns, and in the third case gives a matrix with Subjects as rows and (Contexts x Objects) as columns. Each matrix gives a specific point of view of the data as shown in Figure 2 [1].
We unflattened the tensor along the Context dimension, thus a 2 dimensional matrix of Subject vs Object for each Context was obtained. Each matrix was visulalized with the help of a Self-Organizing Map [5].

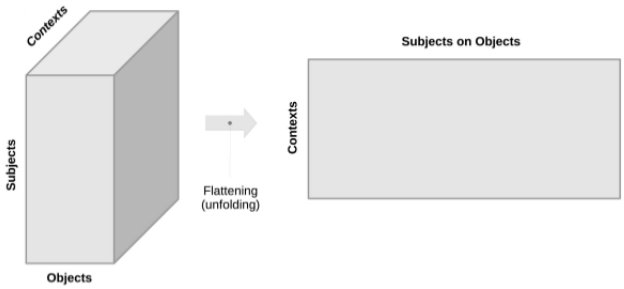


Fig. 2. The $S \times O \times C$-element subjectivity data flattened into a matrix in which each row corresponds to a context and each column to a unique combination of a subject and an object. The number of columns in this matrix is $S \times O$ and the number of rows is C.

3 Implementation

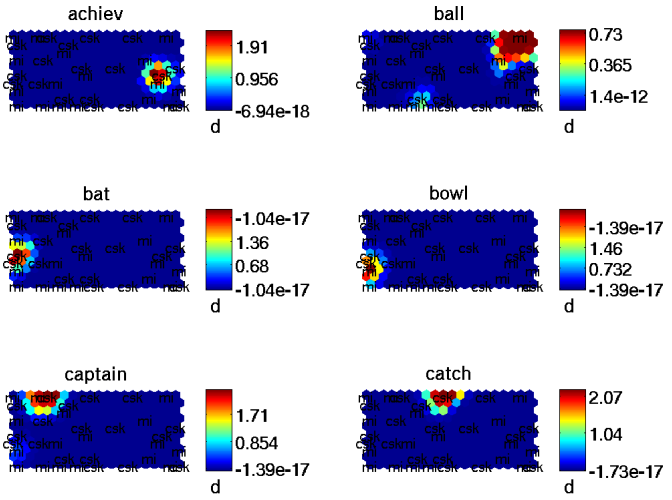
A series of IPL cricket matches [6] between the teams Mumbai Indians (MI), Chennai Super Kings (CSK), Deccan Chargers (DC), Delhi Daredevils (DD), Kings XI Punjab (KXIP), Kolkata Knight Riders (KKR), Pune Warriors India (PWI), Rajasthan Royals (RR) and Royal Challengers Bangalore (RCB) were chosen and 2000 tweets were collected. But matches between MI and CSK was only chosen to process and 494 tweets of both teams were collected. The mining of tweets was implemented by a Java routine, this involves applying a filtration query on the set of tweets for specific teams (subjects). The Twiiter4J [7] Java library was used for this purpose. Filtered tweets were saved in the mySQL database. Then, the occurrence was obtained by reading the tweets from the mySQL database, and most frequently occurring words were taken as Objects. The most frequently occurring words associated with cricket were the contexts and the resultant 3 dimensional and 2 dimensional data was rendered with MATLAB. SOM toolbox [8] was used to analyse the 2 dimensional data.

4 Contribution of team members

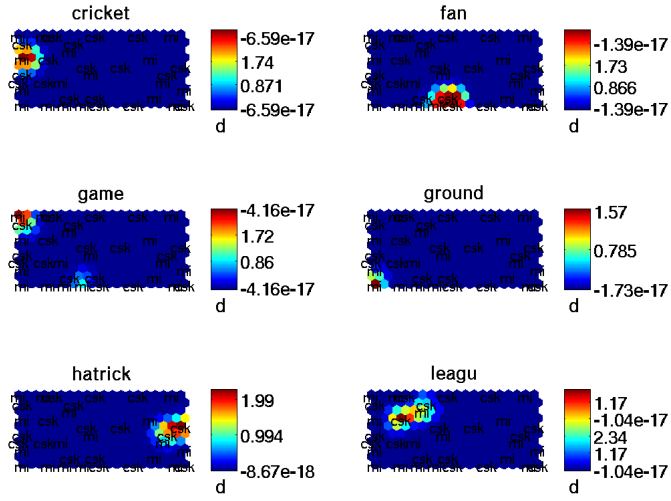
The initial 2000 tweets were collected by Joao. The tweet filtering and text preprocessing and SOM was done by Santosh, Ajay wrote the final report.

5 Results

The figures 3,4 and 5 show the SOM of Subjects X Objects for different contexts (different cricketing terminologies). There are certain contexts that both teams share, like match, over, patnership, achieve, cricket and fan. Each team has its own captain, so we could expect nearly simliar number of tweets for both captains and that is observable in the SOM for context captain. However, there are some other contexts which show that one team was more discussed than the other. If we look at the SOM for catches, CSK team occupies a higher value (brown region) in the SOM compared to MI team (blue region). We could relate this to the fact that in this match there was a brillant catch by a CSK team member, which resulted in greater tweeting. Similarly, the SOM for context win has MI team in slightly higher colour zone than CSK.
We could conclude that MI won the match and if we check the match on which this tweet analysis was done, MI had indeed won the match but by a very narrow

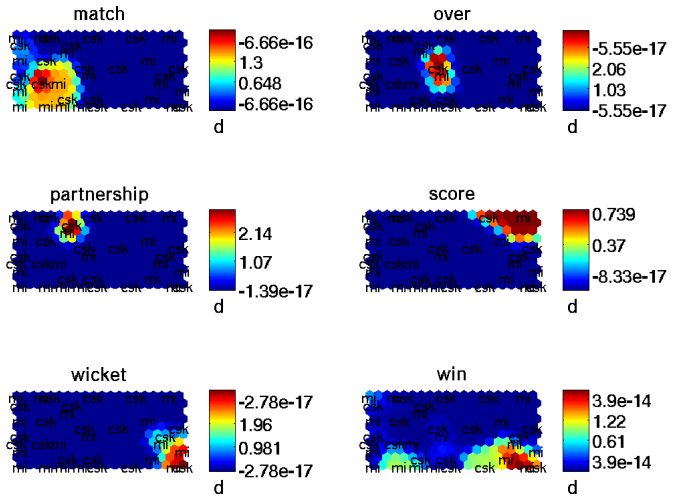


margin. So in the SOM we notice at the right bottom corner, both MI and CSK getting very high distance values but MI is slightly higher. Even the SOM for the context game has a simlar distribution because there could be higher occurrences of phrases such as "MI won the game". Then, the SOM for the context wickets has no distinct separation for either team.



This could be because both teams lost 8 wickets each during the match. However, the SOM for score context shows MI in the top right corner in a distinct zone, maybe highlighting the fact that MI chased a good score.
Another interesting SOM is the context for ground where again MI has a distinct higher value zone, clearly pointing to the fact that the match was played in a ground with more MI supporters, which is indeed the case as the match was played in MI's home stadium, Wankhede.

The U matrix for Context X Object for both Subjects has also been plotted and that is shown in Figure 6 and the U matrix with context labels has been shown in Figure 7. If we superimpose the two, we notice that in the left bottom corner the



Figures 3,4 and 5 show the SOM s for different Contexts

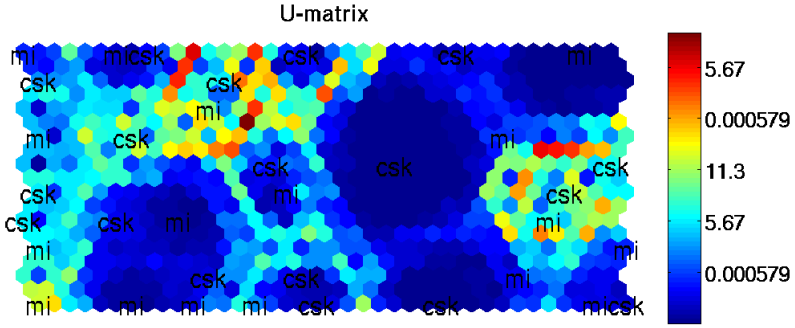


Figure 6 : U matrix for Context X Object for both Subjects

area for MI in Figure 6 and the area occupied by win in Figure 7 match each other perfectly. In the top centre of both figures we have CSK and catch. These results the same what we have obtained with the individual context SOMs.

6 Conclusion

The GICA method when applied to IPL tweets did produce some interesting results with the SOM s for contexts. However, we were guessing as to what the SOM s could be indicating and trying to correlate to the match statistics. Hence, without the match statistics, we cannot infer much from the SOM. Another point that should be kept in mind is that tweets could be positive or negative regarding matching context words. Tweets could also be retweets thus ending up with duplicity. Careful selection of Contexts and Subjects will help improve twitter anaysis using GICA.

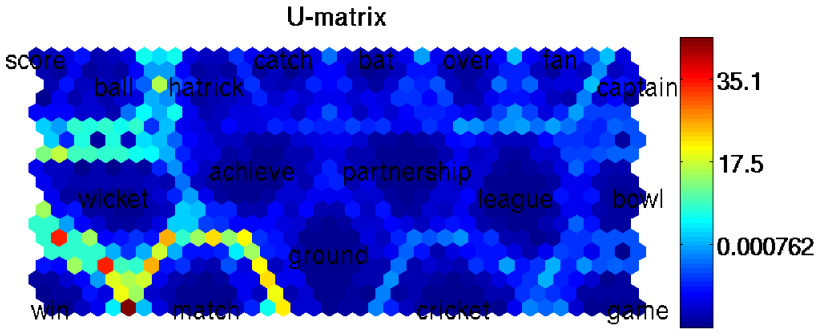


Figure 7 : Context U matrix

References

[1] Honkela et al. Subjects on Objects in Contexts: Using GICA Method to Quantify Epistemological Subjectivity , IEEE World Congress on Computational Intelligence, 2012

[2] Indian Premier League Official Website, http://www.iplt20.com/

[3] Twitter channel of Indian Premier League, https://twitter.com/IPL/

[4] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Review, vol. 51, no. 3, pp. 455500, September 2009

[5] T. Honkela, Self-Organizing Maps in Natural Language Processing, Ph.D. dissertation, Neural Networks Research Centre, Helsinki University of Technology, Espoo, Finland, 1997.

[6] Indian Premier League - 49th match Mumbai Indians v Chennai Super Kings at Mumbai May 6 2012 http://www.espncricinfo.com/indian-premier-league-2012/engine/current/match/548355.html

[7] Twitter 4j library http://twitter4j.org/en/index.html

[8] SOMToolbox, http://www.cis.hut.fi/somtoolbox/