ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 4 April, 2014 Page No. 5335-5337

Mining Hashtags: The Tweet Suggestion System

Anuraag Vikram Kate [1] Nikilav P V[2] Kasthuri Rengan[3]

[1]SASTRA University, School of Computing, Thanjavur, India
anuraag.avk@gmail.com
[2]SASTRA University, School of Computing, Thanjavur, India
nikilav.v@gmail.com
[3]SASTRA University, School of Computing, Thanjavur, India
krengan21@gmail.com

Abstract*: Nowadays social networking has become a requisite and obsession rather than a service. In this ever progressive system, we propose a tweet suggestion system that enhances the basic features of the prevailing tweet structure. In this paper we explore the framework of mining multiple hashtags from tweets in the data repository and suggesting the tweets to the user based on their previous tweets. The algorithm employed here is the Apriori algorithm that mines the frequent data item-set and presents us with the hashtags required for the suggestion.*

Keywords- Hashtags, association rule, apriori, twitter

1. Introduction

Data mining is a dominant knowledge with great prospect to help companies concentrate on the significant information in the data they have gathered about the behaviour of their customers and potential customers. It discerns information within the data that queries and reports can't effectively reveal. The employment of the apriori algorithm effectively produces the candidate sets and thereby gives the frequent item sets. The main dilemma that the users face is what matter to tweet about. The system proposed here suggests the tweets using the hashtags hence putting the users out of their predicament. The usage of database integration is portrayed here to retrieve the tweets dynamically.

2. Problem Assessment

The Micro Blogging and Social Networking era has almost reached its breaking point. People are logging in not knowing what to tweet about and even if they do, it soon turns into something of a tweet war. The problem here is, though there is something called “trending topics”, it is widely general and most of the time, does not cover topics you want to tweet about. For example, during elections in the U.S.A or IPL in India, trending topics are mostly surrounding these high profile events. So, if you are from Afghanistan and want to know what is trending around you, then you are at a loss. Then again it is not realistic to have trending topics based on geographic location, because that defeats the whole purpose of trending topics. A system of prediction is required, where a person’s previous tweet can be used to suggest other topics for him to tweet about.

3.Scope

The scope of this system is highly limited to social networking sites that use hashtags. Twitter, the company that started this trend is the obvious target. But the use of hashtags is also growing in websites like Facebook and Google+. Though it is not useful now, it can be used in the future, if and when hashtags play as big a role there, as they do in Twitter.

4.Methodology

4.1 Data Collection:

A data set of 20 samples has been taken for analysing the proposed methodology. This was a randomly generated dataset which mirrors hashtags used by real users on Twitter. The data was then organized in .arff format. The data set, before and after organization, is shown below.

Tweet No	Hashtags
1	#ICC, #IPL
2	#CSK, #RR, #IPL
3	#CSK, #MSD, #SRINI
4	#BCCI, #SG
5	#ICC, #SRINI, #BIG3
6	#BCCI, #SG, #SC
7	#CSK, #RR
8	#CSK, #RR, #IPL, #BCCI
9	#SRINI, #CSK
10	#MSD, #CSK
11	#IPL, #SG
12	#CSK, #RR, #MSD
13	#CSK, #RR, #IPL
14	#ICC, #BIG3
15	#MSD, #CSK
16	#CSK, #RR
17	#IPL, #BCCI
18	#RR, #IPL
19	#SRINI, #BCCI
20	#ICC, #SRINI

Fig 1: Raw Hashtag Data

```
@relation hashtag

@attribute #ICC {yes,no}
@attribute #IPL {yes,no}
@attribute #CSK {yes,no}
@attribute #RR {yes,no}
@attribute #SRINI {yes, no}
@attribute #BCCI {yes, no}
@attribute #SG {yes, no}
@attribute #SC {yes, no}
@attribute #BIG3 {yes, no}
@attribute #MSD {yes, no}

@data
yes,yes,?,?,?,?,?,?,?,?
?,yes,yes,yes,?,?,?,?,?,?
?,?,yes,?,yes,?,?,?,?,yes
?,?,?,?,?,yes,yes,?,?,?
yes,?,?,?,yes,?,?,?,yes,?
?,?,?,?,?,yes,yes,yes,?,?
?,?,yes,yes,?,?,?,?,?,?
?,yes,yes,yes,?,yes,?,?,?,?
?,?,yes,?,yes,?,?,?,?,?
?,?,yes,?,?,?,?,?,?,yes
?,yes,?,?,?,?,yes,?,?,?
?,?,yes,yes,?,?,?,?,?,yes
?,yes,yes,yes,?,?,?,?,?,?
yes,?,?,?,?,?,?,?,yes,?
?,?,yes,?,?,?,?,?,?,yes
?,?,yes,yes,yes,?,?,?,?,?
?,yes,?,?,?,yes,?,?,?,?
?,yes,?,yes,?,?,?,?,?,?
?,?,?,?,?,yes,yes,?,?,?
yes,?,?,?,yes,?,?,?,?,?
```

Fig 2: Data organized in .arff format

4.2 Data Pre-processing:

Data Pre-processing is an often neglected but important step in the data mining process. Data gathering methods are often loosely controlled, resulting in out-of-range values impossible data combinations, missing values, etc. Analysing data that has not been carefully screened for such problems can procedure misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. The methodology used for data pre-processing is data reduction. Under data reduction we use the data compression technique. We use data compression because the overlay of data present is too excessive for the specific algorithm employed and hence the necessity.

The technique used here is principal component analysis where each hash tag is specifically and singularly analysed. Pre-processing is not required for this sample dataset as it is purely fictional, but for Twitter's local repository that gets millions of tweets every day, it has to be done to ensure speedy results.

4.3 Algorithm:

This system uses apriori algorithm to find frequent items sets to suggest to the user, based on their previous tweets. The algorithm is given in the figure below.

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon)\\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}\\
&\quad k \leftarrow 2\\
&\qquad \textbf{while } L_{k-1} \neq \emptyset\\
&\qquad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}\\
&\qquad\quad \textbf{for } \text{transactions } t \in T\\
&\qquad\qquad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}\\
&\qquad\qquad \textbf{for } \text{candidates } c \in C_t\\
&\qquad\qquad\quad count[c] \leftarrow count[c] + 1\\
&\qquad\quad L_k \leftarrow \{c \mid c \in C_k \wedge\ count[c] \geq \epsilon\}\\
&\qquad\quad k \leftarrow k + 1\\
&\qquad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

Fig 3: The Apriori Algorithm

5. Implementation and Execution

The Weka tool was used to run the Apriori Algorithm with the sample dataset, so as to generate rules and frequent item sets. The results are shown below.

```
Apriori
=======

Minimum support: 0.2 (4 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 3

Best rules found:

 1. #MSD=yes 4 ==> #CSK=yes 4    conf:(1)
 2. #RR=yes 7 ==> #CSK=yes 6    conf:(0.86)
 3. #CSK=yes 10 ==> #RR=yes 6    conf:(0.6)
 4. #RR=yes 7 ==> #IPL=yes 4    conf:(0.57)
 5. #IPL=yes 7 ==> #RR=yes 4    conf:(0.57)
```

Fig 4: Results obtained using Weka 3.6

.

The Weka tool, generated rules for a minimum support of 4 and a minimum confidence of 0.5. However, on a real time basis, if this was to be implemented, the support and confidence has to be increased as the original Twitter data repository will contain a huge number of hashtags even after data reduction, and so more rules can be generated and more specific results can be obtained.

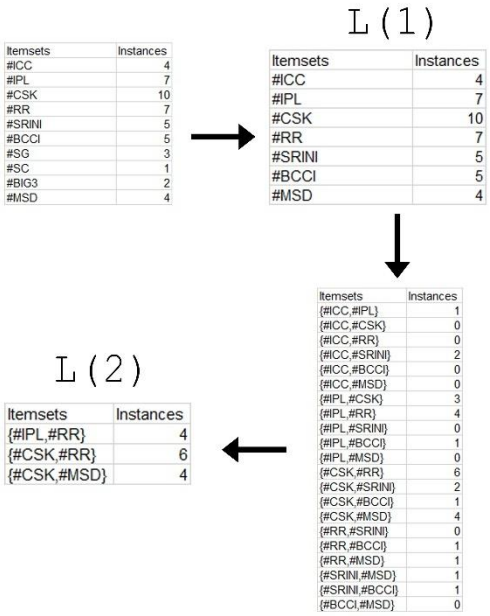

Fig 5: A representation of the working of Apriori algorithm

The above diagram shows the working of the Apriori Algorithm for this dataset. The diagram supports the results presented by Weka.

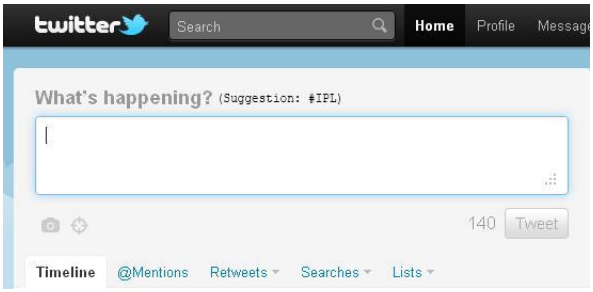

Fig 6: A representation of Twitter after implementing this system.

6. Conclusion

Thus based on the above findings we conclude that the modifications presented by our framework clearly enhance the viability and the efficiency of the existing tweet system. This system makes the tweeting structure more comprehensible and compact. It also add to the betterment in the organisation of the tweet hashtags in the data repository. This mining mechanism is an added feature to the existing tweet structure that makes the lives of the tweeters easy and in this comfort-driven 21st Century, easy is the name of the game.

7. References

[1]Usama Fayad, Gregory Piatesky-Shapiro, Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 2014.
[2] "Special Issues on Data Mining",CACM,1996
[3]Ronny Kohavi, Foster Provost, "Application of Data Mining to Electronic Commerce", Kluwer Academic Publishers, 2001.
[4]"Special Issue on Data Mining", IEEE Computer, 1999.
[5]S. G. Esparza, M. P. O'Mahony and B. Smyth, "Towards Tagging and Categorisation for Micro-blogs", AICS'10,2010.
[6] Allie Mazzia, James Juett, "Suggesting Hashtags on Twitter", University of Michigan, 2013.
[7] W. Wu, B. Zhang and M. Ostendorf,"Automatic generation of personalized annotation tags for Twitter users", HLT'10, 20