Twitter Association Rule Mining using Clustering and Graph Databases

Alessandro Campi
Politecnico di Milano, Dipartimento di Elettronica e Informazione, piazza Leonardo da Vinci 32, Milano, Italy

Corrado Palese
Politecnico di Milano, Dipartimento di Elettronica e Informazione, piazza Leonardo da Vinci 32, Milano, Italy

ABSTRACT

In this scenario, the need to efficiently analyze this kind of data is increasing because of characteristics of such big data, especially their huge and sometimes unpredictable variety. Twitter alone, with 320 M active users every month and more than 500 M tweets per day, could represent an important source of information. For this research, we are focusing solely on social networks. The reason for this choice is that they are increasingly becoming a platform where people will comfortably update their status and share or retrieve information about the world in real time. Often news is spreading through them faster than in traditional channels because user capillarity worldwide makes it possible. In particular, we will focus on Twitter, because its micro-blogging nature makes it suitable for this kind of purpose. It questions the concept of a small private community of friends in favor of less private, less personal broadcast communications of common interest. Another reason why we chose Twitter is because semantic value of hashtags, their power in summarizing tweet content and the spreading model through the social network that allows us to highlight clusters of topics by focusing on these tags.One of the objectives of this thesis is to show how data mining can provide useful techniques to deal with these huge datasets for retrieving information to detect and analyze trending topics and the corresponding user's interactions with them. We identified in Association Rules identification and evolution in time, a systematic approach to conduct the analysis.

CCS CONCEPTS

• Information systems; **• Data analytics**; **• Computer systems organization**; **• Real-time system**; **• Real-time system architecture**;

KEYWORDS

Clustering, Association rules, Social media

ACM Reference Format:
Alessandro Campi and Corrado Palese. 2021. Twitter Association Rule Mining using Clustering and Graph Databases. In *2021 the 5th International Conference on Information System and Data Mining (ICISDM 2021), May 27–29, 2021, Silicon Valley, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3471287.3471309

ICISDM 2021, May 27–29, 2021, Silicon Valley, CA, USA

ACM ISBN 978-1-4503-8954-9/21/05…$15.00
https://doi.org/10.1145/3471287.3471309

1 INTRODUCTION

Interest in social networks analysis has been growing recently since they represent one of the biggest resources for the well-known big data context. The main innovation coming from them, is related to the capability of providing a channel of data generation used by billions of users everyday [21]. However, data does not mean information. Data is raw and it should be processed to gain useful and human understandable content.

When we deal with social data, some considerations are needed because as well as being a very big source, it is non-structured, dynamic and normally contains noise. Moreover, the information we can extract from them is usually related to real life events. This means that the analysis should be done in real time to be considered reasonable.

The scientific community seems to accept data mining as the most appropriate technique for generating useful information from such big dataset but a preprocessing phase is normally needed in a non-structured context.

For these reasons our research will focus on social network data provided by Twitter with the aim of finding a systematic way to make the analysis efficient and scalable. We find in association rule mining the best approach to extract useful information in order to identify and track events in real time. We took inspiration from the market basket analysis [2] where the method was used to find correlations between items bought together a certain number of time.

We transpose the idea of transactions into the social context transforming tweets into itemsets composed of the hashtags they contain. We take advantage of the semantic values of hashtags in summarizing tweet content through keywords in order to avoid any text mining phase in preprocessing. In particular, we will examine two key points in our work that makes the whole analysis interesting in terms of time and computational resources.

One is the implementation choice of using graph databases to persist data pulled from the network. This new technology greatly fit the need to model the real-world entity-relationship mechanism especially into a relationships-centered environment, as social networks are.

They allow to explicitly save relationships and to query them without executing join operations that compromise performance when they deal with huge datasets.

The other key point concerns the methodology we use to choose the datasets on which we generate association rules. We do not apply any filters, except from the language, in pulling the network but we developed a method to cluster tweets before executing the association rules mining. The clustering process drastically reduces the attribute space made by hashtags in every single execution of rules mining with the effect of reducing execution time and computational resources needed, without losing interesting information.

2 RULE MINING

In the recent years there have been a lot of researches conducted on Twitter data for trending topic detection and events tracking but none of those were used to track different events simultaneously without applying filters in data collection phase. With our approach, we developed a method for real time clustering of tweets on which we execute the association rules mining analysis to track events in a reasonable time. The clustering process reduces the number of tweets we analyze together, finding which of them are correlated and could generate interesting rules. Dealing with social networks, with the huge number of feeds published daily and with the fact that we decided to not do any pre-filtering activity, we will demonstrate how the clustering process allows to save time during the execution and computational resources without losing information. Finally, once tweets are divided into clusters, the same analysis could be executed on all of them in parallel with a further reduction in overall execution time. The main idea for creating clusters is to find which are the most popular hashtags in the dataset and then find which hashtags are correlating to them. Despite of many researches on this topic, our work shows that any filtering is needed for classifying tweets. Moreover, our algorithm does the classification in real time, explicitly saving a relationship entity between hashtags that appear together when a tweet is pulled from the network, or incrementing the relationship attribute called counter if it already exists.

We will use a graph database to save tweets and relationships between them. Saving relationships explicitly give us a strong competitive advantage with respect to other implementations that use relational databases since we do not need to execute join operations. Finally, in order to find association rules that represents the final objective of our work, we decided to treat tweets as transactions. We took inspirations from association rule mining in market baskets (Agrawal, Imieliński et al. 1993a).

Moreover, we decide to use hashtags as items and tweets containing hashtags as transactions on which applying Weka implementation of the Apriori algorithm.

We define two type of entities:

1. TWEET in which we saved as attributes:

- ID: long - the unique ID Twitter assigns to each tweet.
- Text: string - tweet content.
- Author: long - the author ID Twitter assigns to each user.
- IsRetweet: boolean – true if the tweet is a retweet, false in all other cases.
- IsAnswer: boolean – true if the tweet is an answer to another one, false in all other cases.
- Hashtags: string – comma separated list of hashtags contained by the tweet.
- Timestamp: long – creation timestamp.

2. HASHTAG in which we saved as attributes:

- Text: string - the hashtag itself.

Moreover, we defined two type of relationship: TAGS: unidirectional relationship between a hashtag and a tweet that contains it and APPEAR_TOGETHER: conceptually bidirectional relationship between hashtags that appear together at least in one tweet. Within APPER_TOGETHER relationship we keep track of how many times two hashtags appear together through an attribute called count.

3 ALGORITHM WORKFLOW

The algorithm can be divided into three main phase. The first step is data retrieval in which we extract tweets published in the network and we saved them into the graph.

The second step, is a key point of our research, because doing the clustering we drastically reduce time in finding association rules without losing information. This steps starts specifying one or more hashtags on which we would like to focus on. Once these are defined, all correlated hashtags are extracted, forming the space field in which rules will be calculated. From these hashtags, we then extract all tweets that contains at least one of them. They represent the tweet subset identified, correlated with the topic described by hashtags correlation. At this point, starting from tweets, another step would be needed to find hashtags contained by each of them. This operation implies a high cost in terms of time since we need to follow all incoming relationships for each tweet that belongs to the cluster. In order to avoid this, we add the list of hashtags also as a tweet property in form of string, made by comma separated hashtags. Moreover, it is exactly what we will save in the CSV file and it will represent the corresponding tweet in association rules mining analysis.

The first step of our Java application is the information retrieval. What we consider as information, are tweets that, for a certain period, are published on the social network.

For this research, we decide to not do any prefiltering activities on tweets retrieval except for the language. The choice of filtering on English language is purely technical: give us the opportunity to not consider languages that is not understandable from our side even if, at the same time, represent huge quantity of information on Twitter network (Eg. Chinese, Arab languages, etc).

We pull Twitter network through Twitter4j API, using Twitter streaming endpoint. The choice of using this endpoint is because we are interested in retrieving information that is flowing on the network in a certain period. We do not interest in doing searches on a certain topic but we only want information on which we can find, by our approach, the most trending topics for that time. In order to build an access point to Twitter streaming APIs, all we need is a class that implements a StatusListener. Twitter4j will then create a thread to consume the stream. The mechanism used is an event handling process. Twitter4j implements on the StatusListener class, the method onStatus where we can specify what happens each time we receive a feed. The onStatus function is executed every time a tweet is caught from the network, it is a callback function that return a Status object 10. The Status object contains all the information about the tweet that represent and provides all the functions to access information we need. Every time a tweet is caught, information should be saved in the graph. On the callback execution, a node in the graph of TWEET type is created and relative information is stored. Moreover, every hashtag contained in the feed should be saved as well as a HASHTAG node type, only if it does not exist yet. A relationship of TAGS type should be created to link the hashtag with the current tweet. Finally, all hashtags within the same tweet should be linked each other with an APPEAR_TOGETHER relationship in a full connected subgraph. Information coming from the APPEAR TOGETHER relationships is redundant. As we have

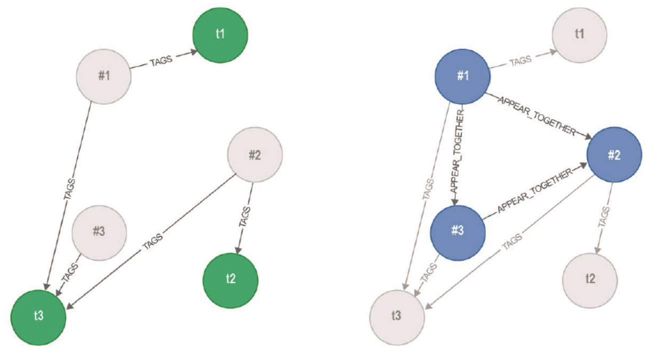

Figure 1: Appear_together relationships

already discussed, even if relationships in Neo4j are declared as monodirectional, they could be seen conceptually bidirectional. Moreover, the Cypher Query Language support the need to cross them in both ways.
Having a look at the figure 1, we can see that we could retrieve hashtags that appear together within a tweet starting from the tweet itself. In fact, each tweet is linked by a TAGS relationship to all hashtags contained. This implies that, starting from a hashtag, we can find all hashtags that appear together with it, passing through tweets that tags. The complexity of the process would be very high since we would need a number of queries equals to the number of tweet containing the hashtag on which we start the correlation analysis. It would represent a serious problem in a real-time process that potentially deal with millions of instances. Instead, the choice of storing APPEAR TOGETHER relationships, even introducing information redundancy, allow us to find out all correlated hashtags just through one query, starting from one of them without depending on the number of instances.

The clustering process begins when hashtags and mutual APPEAR_TOGETHER relationships are stored into the graph. Our algorithm builds cluster in real time when tweets are caught from the network without any additional computation needed. Since our analysis potentially involves billions of tweets and hashtags, we have specified a method to choose which hashtags, within a cluster, should be considered interesting.

First, once data are collected, our algorithm allow us to investigate most popular hashtags on which extract the relative cluster. Otherwise we can also specify one, making a sort of filtering around an interesting topic represented by a hashtag. This scenario is implemented in case we want to analyze a determinate event eventually using the official hashtag. Once one or more hashtags are chosen, the algorithm extract which hashtags in relative subgraphs are considered interesting.

The main idea of filtering hashtags we do not consider interesting is based on the fact that, reducing the space in which we will look for association rules, we obtain a significant decrease in execution time without losing information. We consider a correlation between two hashtags “interesting”, if they appear together a certain number of time in tweets that belong to the sample on which we are doing

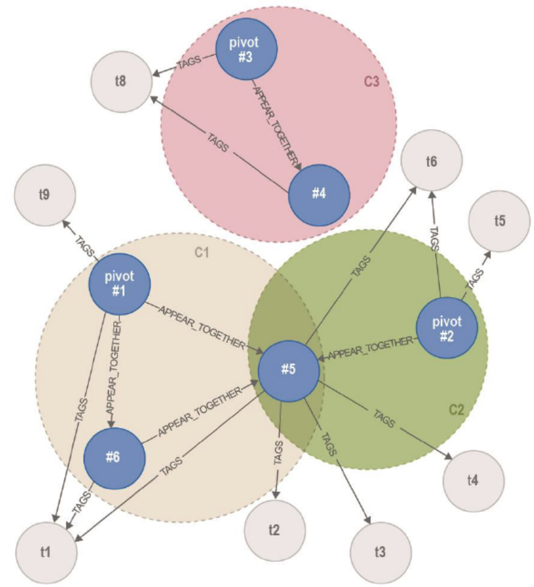

Figure 2: Clustering process

the analysis. To do this systematically, we need to define a threshold value.

In order to explain the mining of the threshold we are going to use an example. From now on, we will call the hashtag chosen to build the cluster, pivot. To build the cluster we only consider hashtags directly connected with the pivot and eventually relationships between them but only if, singularly, all of them reach the threshold value on the direct relationship with the pivot.
To keep track of the number of time two hashtags appear together, we use the count property defined into the APPEAR_TOGETHER relationship. Consider two hashtags: a and b, both connected with the pivot and a threshold s. They will be part of the cluster if and only if the number of time they appear together with the pivot out of the total number of tweets that contain the pivot itself, overcomes s. Moreover, if both satisfy the threshold and they are also mutual connected, this relationship is carried into the cluster. The choice

of using a threshold value on the pre-processing phase seems to be useless since the minimum support applied by association rules mining algorithm will exclude all correlated hashtags from generated rules if they do not overcome the threshold. The reason why we decide to apply a pre-processing filter on hashtags occurrences comes from the need to make the space search smaller, in order to save computational resources. In fact, experimentally, we notice that, without applying this threshold, we generated a search field made by thousands of hashtags. Thinking about how the Apriori algorithm works, most of them, would disappear automatically in frequent itemsets generation but they could cause the generation of a huge number of candidate sets. The very low minimum support we use to calculate association rules do not help the process in this sense. For example, if there are 10^3 frequent 1-itemsets, the Apriori will need to generate more than 10^5 candidate 2-itemset, causing several Weka crashes for RAM lack.

Our choice to adapt tweets as transactions in a transactional database comes from the fact that this kind of data model fits well the need to find association rules (Kotsiantis, Kanellopoulos 2006b). We transform a tweet in a list of item that sum up the content in terms of topics it contains. Scientific community recognizes the semantic value of hashtags in summarizing tweet content (Romero, Meeder et al. 2011) and we exploit it in order to avoid text mining technics to tokenize the tweet. Items that make a transaction representation of the tweet are hashtags it contains. The graph structure allows us to extract hashtags contained by a tweet but, for performance reasons, we introduced a redundancy in data, saving hashtags as a property of the tweet. In this way, every node of TWEET type has the hashtags property: a string with comma separated hashtags text.

The output of this phase is a CSV file we will use as input to run for Weka in order to execute the association rule mining analysis through the Apriori algorithm.

3 EXPERIMENTS

We made different tests, but for lack of space we describe only a test made by the public dataset of Roi Olimpics Tweet. In our test, we define a correlation threshold for hashtags of 2%. In other words, it means we consider two hashtags correlated if they appear together at least the 2% of times the most popular once appears into tweets. For example, if we consider two hashtags, A and B, with A the most popular of them, we say that they are correlated and belong to the same cluster, if at least the 2% of tweets that contains A, contains also B. Moreover, we define the support and confidence threshold in association rules mining as follow. For confidence, we use the value of 0.5% because we noticed that, considering the first 100 rules in support order for every cluster, they all overcome this percentage. For support, we did some experiment in order to find the best threshold to make the analysis reasonable. We notice that we cannot find almost any association rules with a less than 0.001% minimum support. Even if this value could seem too low, it is justified by the fact that we are working on a very huge sample without doing any pre-filtering activity. Values are in line with other researches conducted on the same field (Adedoyin-Olowe 2015). We need to mentioned that, when we run the Apriori algorithm without using clustering, we use the same confidence but a support of 0.0001% since we analyze ten clusters all together that approximately means ten times the number of tweets with respect to only one cluster.

The first test has been done on tweets retrieved from the 07th of August to the 13th of August 2016.

- Sample dimension: 490971 tweets

- Execution time in association rules mining on clusters built around ten most popular hashtags:

- MTVHottest – 0 min 1.407 sec
- PushAwardsLizQueen – 0 min 1.056 sec
- DolceAmoreOperation1010 – 0 min 1.096 sec
- VeranoMTV2016 – 0 min 1.400 sec
- Rio2016 – 0 min 4.021 sec
- PushAwardsKathNiels – 0 min 1.021 sec
- ALDUBsaAfrica – 0 min 1.082 sec
- MUFC – 0 min 1.120 sec
- Gameinsight – 0 min 1.042 sec
- USA – 0 min 6.625 sec

Execution time in association rules mining without using clustering: 2 min 28.89 sec.

Full training set (hashtags) made by 86 attributes: #Archery, #ARG, #ArtisticGymnastics, #Athletics, #AUS, #badminton, #basketball, #Basketball, #bbcrio2016, #BRA, #BREAKING, #BringOnTheGreat, #Bronze, #bronze, #CAN, #CHN, #cycling, #CyclingRoad, #CyclingTrack, #DEN, #DipaKarmakar, #ESP, #Fencing, #fencing, #FIJ, #Fiji, #FinalFive, #football, #FRA, #GBR, #GER, #Gold, #GOLD, #gold, #gymnastics, #Hockey, #Ind, #IND, #IRQ, #ITA, #JAM, #JPN, #Judo, #KheloIndia, #KOR, #LalitaBabar, #MichaelPhelps, #NGR, #NZL, #Olympic, #OlympicGames, #olympics, #Olympics, #Olympics2016, #OpeningCeremony, #Phelps, #PHI, #POR, #Rio, #Rio2016, #RioOlympics, #RioOlympics2016, #rowing, #RSA, #Rugby7s, #RugbySevens, #Silver, #silver, #SILVER, #SimoneBiles, #SWE, #Swimming, #swimming, #TableTennis, #TeamCanada, #TeamGB, #TeamUSA, #Tennis, #tennis, #USA, #USABMNT, #USABWNT, #volleyball, #weightlifting, #WirfuerD, #yourteam

Whit the minimum support of 0.001% and a minimum confidence of 0.5%, 115 rules have been found. For reasons of clarity, we only show the first ten rules, ordered by confidence, and some others that allow to understand the power of our algorithm to highlight events and tracks their outcomes. Other rules have been omitted in this paragraph because they do not give any information because of the absence of semantic values in hashtags or they are redundant with rules presented.

1. yourteam=t 122 ==> Rio2016=t 122 <conf:(1)>
2. Swimming=t BringOnTheGreat=t 46 ==> Rio2016=t 46 <conf:(1)>
3. KheloIndia=t Rio2016=t 28 ==> Hockey=t 28 <conf:(1)>
4. Gold=t ArtisticGymnastics=t 90 ==> USA=t 89 <conf:(0.99)>
5. CHN=t ITA=t 35 ==> USA=t 34 <conf:(0.97)>
6. CHN=t Rio2016=t ITA=t 33 ==> USA=t 32 <conf:(0.97)>
7. CHN=t USA=t 48 ==> Rio2016=t 46 <conf:(0.96)>
8. bronze=t gold=t 36 ==> silver=t 34 <conf:(0.94)>
9. CHN=t ITA=t 35 ==> Rio2016=t 33 <conf:(0.94)>
10. CHN=t USA=t ITA=t 34 ==> Rio2016=t 32 conf:(0.94)
11. Olympics=t NGR=t football=t 29 ==> Rio2016=t 27 conf:(0.93)
16. KheloIndia=t Hockey=t 31 ==> Rio2016=t 28 conf:(0.9)
17. GOLD=t FIJ=t 30 ==> Rio2016=t 27 <conf:(0.9)>

18. NGR=t football=t 38 ==> Rio2016=t 34 conf:(0.89)
25. NGR=t DEN=t 35 ==> Rio2016=t 30 <conf:(0.86)>
43. Rugby7s=t Fiji=t 42 ==> Rio2016=t 31 conf:(0.74)
45. ArtisticGymnastics=t USA=t 123 ==> Gold=t 89
53. Gold=t Rugby7s=t 46 ==> FIJ=t 32
66. Fiji=t Rio2016=t 49 ==> Rugby7s=t 31 conf:(0.63)
77. Gold=t FIJ=t 52 ==> Rugby7s=t 32 <conf:(0.62)>
78. Tennis=t 149 ==> Rio2016=t 91 conf:(0.61)

An interesting information comes from rules 3 and 4. In the first one, we immediately see that one of the most trending argument is related to hockey discipline and the Indian team. During these days, in fact, the men Indian team reached the quarter-finals after thirty-six years. From these rules, we focus on the event but we do not have any suggestion about the outcome. In the second one, instead, we clearly understand that USA wins the gold medal in artistic gymnastic discipline. In this case, we do not just have information about the event itself, but also about the outcomes. In fact, having a look at the training set, we find the hashtag #SimoneBiles, the American athlete who won four gold medals in four different artistic gymnastic disciplines exactly in these days. Probably, with a lower minimum confidence index, we would find same rules involving also the SimoneBiles hashtag that it would give us a complete information. From rules 11, 18 and 25, we can see that something happens that involves Nigerian and Danish football teams. In fact, on the 13th of August, the Nigerian football team won against Denmark and reached the semi-finals. Finally, from rules 17, 43, 57, 77, we can see that Fiji rugby team won the gold medal. Moreover, it was the first-ever Olympic medal for the pacific island nation. One more interesting thing is that, even with only a very small training test, we have been able to underline the most important events happened during days in which we made the analysis.

4 CONCLUSIONS

The main challenge of this work was to find out a systematic and scalable approach to deal with non-structured data analysis. Our decisions to focus on Twitter data because of its unstructured nature and the huge number of raw data it can provide, suggests to us the use of data mining techniques as a possible way to conduct the analysis. We show as the direct application of one of the most famous association rule mining algorithm, the Apriori, is not possible without a structuring phase in preprocessing. At this point we took inspiration from market basket analysis, transforming tweets in itemsets made by hashtags they contain, building a sort of transactional database. The objective was to highlight, into the Twitter feed, interesting topic and eventual correlations between them, studying the correlation between hashtags. Trending topics in social networks are usually related to real-life events and the user's engagement towards them. Given the consideration above, we deal with some additional constraints. For example, the fact that real-life event detection should be done in real time to be considered valuable. For this reason, we focused our research on finding a way to reduce execution time and computational resources needed.

We supposed that the main problem in executing association rules mining could be the attribute space analyze each time and not the number of transactions.

Other works on this field apply a prefiltration phase in pulling the network to limit the number of keywords but we only filtered non-english tweets. To solve this issue, we proposed a clustering process to split hashtags, assigning them to smaller macro-areas before looking for rules. Results obtained confirms our hypothesis, showing as the execution time is in a quadratic relation with respect to the number of attributes. The correctness of the process is confirmed to the fact that we do not lose information in terms of rules, executing the same analysis without applying clustering on the same dataset.

REFERENCES

[1] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.
[2] AGRAWAL, R., IMIELIŃSKI, T. and SWAMI, A., 1993a. Mining association rules between sets of items in large databases, Acm sigmod record 1993a, ACM, pp. 207-216.
[3] AGRAWAL, R. and SRIKANT, R., 1994. Fast algorithms for mining association rules, Proc. 20th int. conf. very large data bases, VLDB 1994, pp. 487-499.
[4] ALONSO, O., CARSON, C., GERSTER, D., JI, X. and NABAR, S.U., 2010. Detecting uninteresting content in text streams, SIGIR Crowdsourcing for Search Evaluation Workshop 2010.
[5] ASUR, S. and HUBERMAN, B.A., 2010. Predicting the future with social media, Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM Int. Conference on 2010, IEEE, pp. 492-499.
[6] BALMIN, A., HRISTIDIS, V. and PAPAKONSTANTINOU, Y., 2004. Objectrank: Authority-based keyword search in databases, Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 2004, VLDB Endowment, pp. 564-575.
[7] BECKER, H., NAAMAN, M. and GRAVANO, L., 2011. Beyond trending topics: Real-world event identification on twitter.
[8] CORNEY, D., MARTIN, C. and GÖKER, A., 2014. Spot the ball: Detecting sports events on Twitter. Advances in information retrieval. Springer, pp. 449-454.
[9] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.FENG, W. and WANG, J., 2014. We can learn your# hashtags: Connecting tweets to explicit topics, 2014 IEEE 30th International Conference on Data Engineering 2014, IEEE, pp. 856-867.
[10] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.FRANK, E., HALL, M., HOLMES, G., KIRKBY, R., PFAHRINGER, B., WITTEN, I.H. and TRIGG, L., 2005. Weka. Data Mining and Knowledge Discovery Handbook. Springer, pp. 1305-1314.
[11] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.HAN, J., PEI, J. and YIN, Y., 2000. Mining frequent patterns without candidate generation, ACM Sigmod Record 2000, ACM, pp. 1-12.
[12] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.JAVA, A., SONG, X., FININ, T. and TSENG, B., 2007. Why we twitter: understanding microblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis 2007, ACM, pp. 56-65.
[13] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.KOTSIANTIS, S. and KANELLOPOULOS, D., 2006a. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering, 32(1), pp. 71-82.
[14] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.KWAK, H., LEE, C., PARK, H. and MOON, S., 2010. What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web 2010, ACM, pp. 591-600.
[15] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.MATHIOUDAKIS, M. and KOUDAS, N., 2010. Twittermonitor: trend detection over the twitter stream, Proceedings of the 2010 ACM SIGMOD International Conference on Management of data 2010, ACM, pp. 1155-1158.
[16] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.MONTAG, D., 2013. Understanding neo4j scalability. White Paper, Neotechnology, MOONEY, R.J. and BUNESCU, R., 2005. Mining knowledge from text using information extraction. ACM SIGKDD explorations newsletter, 7(1), pp. 3-10.
[17] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.MORSTATTER, F., PFEFFER, J., LIU, H. and CARLEY, K.M., 2013. Is the sample good enough? comparing

data from twitter's streaming api with twitter's firehose. arXiv preprint arXiv: 1306.5204, .
[18] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.NAAMAN, M., BOASE, J. and LAI, C., 2010. Is it really about me?: message content in social awareness streams, Proceedings of the 2010 ACM conference on Computer supported cooperative work 2010, ACM, pp. 189-192.
[19] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.ROBINSON, I., WEBBER, J. and EIFREM, E., 2015. Graph Databases: New Opportunities for Connected Data. "O'Reilly Media, Inc.".
[20] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.ROMERO, D.M., MEEDER, B. and KLEINBERG, J., 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter, Proceedings of the 20th international conference on World wide web 2011, ACM, pp. 695-704.
[21] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.WU, X., ZHU, X., WU, G. and DING, W., 2014. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), pp. 97-107.
[22] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.YAMAGUCHI, Y., TAKAHASHI, T., AMAGASA, T. and KITAGAWA, H., 2010. Turank: Twitter user ranking based on user-tweet graph analysis. Web Information Systems Engineering–WISE 2010. Springer, pp. 240-253.
[23] ADEDOYIN-OLOWE, M., 2015. An association rule dynamics and classification approach to event detection and tracking in Twitter.YANG, M., LEE, J., LEE, S. and RIM, H., 2012. Finding interesting posts in twitter based on retweet graph analysis, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval 2012, ACM, pp. 1073-1074.