iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VI **Month of publication:** June 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Online News Trend Discovery Using Apriori Algorithm

Navneet Soni[1], Dr. S M Ghosh[2]
M. Tech Scholar, Associate Professor
Rungta College of Engineering and Technology, Bhilai

Abstract— The tremendous growth of Internet has given rise to many news article websites. These news article websites maintain a large amount of information. Discovering usage pattern has direct implication in improving website design, generating advertisements and content distribution. Evaluating popularity prior to release is somewhat desirable, allowing the possibility of reworking the articles and changing the manner of their publication. This task of pattern discovery comes under the broad area of web usage mining. There are a lot of techniques/algorithms that can be employed for web usage mining. In this paper, we discuss the use of Apriori algorithm to identify usage pattern by establishing relationships among various attributes related with a given URL. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. During this project, we have used association rule mining (Apriori algorithm) to try to discover pattern of popularity on the dataset published by Mashable[1], summarizing a heterogeneous set of features about articles published by it in a period of two years.
Keywords- Pattern discovery; Association rule; Frequent Item set; Apriori.

I. INTRODUCTION

Online news article is an engaging activity enjoyed by huge number of mobile users. Due to advent in wireless mobile internet connectivity, online news consumption have grown tremendously. For a news article to be popular, it is essential for it to propagate to a large number of users within a short time. News articles are extremely time sensitive by nature. What influence does the time and genre of the article have on popularity? Due to the time-sensitive aspect and fierce competition for attention, accurately estimating the extent to which a news article will spread on the web is quite valuable to journalists, advertisers, content providers and news recommender systems. Also politicians and activists are using web as the platform in influencing public opinion. However, predicting online popularity pattern of news articles is a challenging task as it is influenced by a number of complex factors like subject sensitivity, content of the article, geography or temporality etc.

Most early researches have focused on determining popularity prior to actual release. Kirutikha M, Jadhav R [1] found strong association rules among related web pages sampled from web Server logs. The support and confidence values of extracted rules were considered for obtaining the interest of the web visitors. Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Serge Fdida [2] proposed a linear log popularity prediction model as an effective solution to online news ranking, with a performance that can evenly match more customized learning to rank algorithms. Bandari R, Asur S, Huberman et. al. [3] developed their own t-density factor to be used with regression and classification to predict ranges of popularity on Twitter [2] with an overall accuracy of 84%. A growing number of recent studies predict spread of information based on early measurements(using early votes on Digg[3], likes on Facebook[4], click-thoughts, and comments on forums and sites) Szabo and Huberman 2010 [4] found eventual popularity of items posted on Youtube [5]and Digg has a correlation with their early popularity.

One of the most useful aspect of the data mining is not the prediction itself, but the insights gain from algortithm and the structural pattern that algorithms can reveal. This is not the first time that data mining techniques are being applied for prediction and cultural artifacts. Several companies have applied data mining to music and movies over the past years. In the following section, we have

used association rule mining (Apriori algorithm) to try to discover pattern of popularity on the dataset published by Mashable, summarizing a heterogeneous set of features about articles published by it in a period of two years.

II. PROBLEM DEFINITION

Mining web log for relationship among links provides a pattern of user behavior as how a user navigate among different pages, but it overlooks what is 'inside' a link as these web logs do not usually maintains information about what is inside. This kind of relationship among URLs fails to capture the content's genre and the subject of matter. Whereas, other researchers, who have made social networking as a tool for measuring the interestingness of a particular article, often ignores the pattern of relationship among the links or network topology.

III. DATA SET

The articles were published by Mashable. This dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided URLs. The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method.

The dataset used here has a lot of attributes and sufficient number of instances that can fit our purpose. It is expected to derive a number of interesting relationships among a number of attributes in our quest for pattern discovery and ultimately using the knowledge obtained from the patterns discovered in effectively predicting the popularity of future articles.

Number of Attributes: 60 (58 predictive attributes, 2 non-predictive)

IV. PREDICTION METHOD

The relationships among links as well as the interestingness of the content both can be accounted through association rules. Association rule learning is a method for discovering interesting relations between variables in large databases. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." It is intended to identify strong rules discovered in databases using different measures of interestingness. Association rules are created by analyzing data for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true. We are using Weka tool version 3.7.4 for the same.

A. Pre-Processing And Dimension Reduction

There aren't any missing values in the dataset and all the values of the attributes are sufficiently cleansed. However, since Apriori works on nominal attribute values, all the numeric attributes values in the data set are converted into nominal values using appropriate filters.

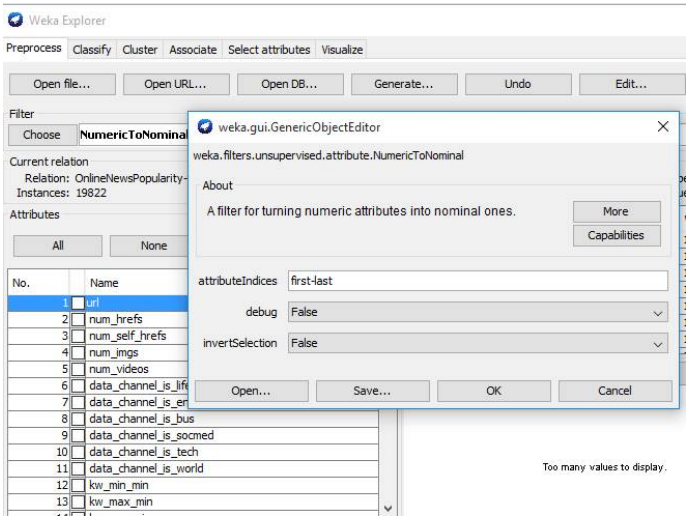


Figure 1: Numeric To Nominal filter in Weka.

There are a lot of attributes in this dataset and few of them are not relevant for our study; therefore some dimensions would be better removed. For our study we have only included attributes that tells the day of publication and the genre of the articles. However since

data mining is not a rule fixed set of task and mostly uncertain , so these dimensions could be added (if needed) for a different insight altogether. In fact, we will try a different set of mix attributes to gain different kinds of relationships.

1) *Finding frequent patterns using Apriori algorithm:* **Apriori** is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Key concepts of Apriori algorithm are:
   a) *Frequent Item sets:* The sets of item which has minimum support (denoted by L_i for i^{th}- Item set).
   b) *Apriori Property:* Any subset of frequent item set must be frequent.
   c) *Join Operation:* To find L_k , a set of candidate k-item sets is generated by joining L_k-1 with itself.

2) *Association rule mining is a two-step process:*
   a) Find the frequent item sets, i.e., the sets of items that have at least the minimum support σ.
   b) Using the frequent item sets to generate (strong) association rules that satisfy the minimum support and minimum confidence γ.

Using the Apriori algorithm we want to find the association rules that have minSupport=65% and minimum confidence=80%. We will do this using WEKA GUI. For better comprehension we are eliminating Lift, Leverage and conviction values from the results.

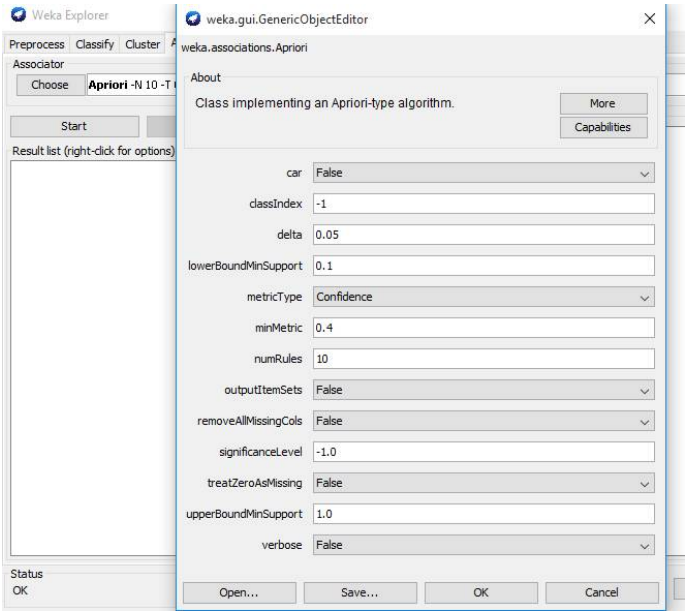


Figure 2: Parameter values for our test.

V. RESULT

Here is the screenshot of our test run:

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.8 -D 0.05 -U 0.7 -M 0.1 -S -1.0 -c -1
Relation: OnlineNewsPopularity-weka.filters.unsupervised.instance.RemovePercentage-P50.0-weka.filters.unsupervised.attribute.Remove-R1-13,20-31,40-61-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 19822
Attributes: 14
data_channel_is_lifestyle
data_channel_is_entertainment

data_channel_is_bus
data_channel_is_socmed
data_channel_is_tech
data_channel_is_world
weekday_is_monday
weekday_is_tuesday
weekday_is_wednesday
weekday_is_thursday
weekday_is_friday
weekday_is_saturday
weekday_is_sunday
is_weekend

=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.65 (12884 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 7
Generated sets of large itemsets:
Size of set of large itemsets L(1): 14
Size of set of large itemsets L(2): 75
Size of set of large itemsets L(3): 123
Size of set of large itemsets L(4): 81
Size of set of large itemsets L(5): 26
Size of set of large itemsets L(6): 3
Best rules found:

A. *data_channel_is_lifestyle=0 13863 ==> weekday_is_saturday=0 13863 conf:(1)*
B. *data_channel_is_lifestyle=0 13863 ==> weekday_is_sunday=0 13863 <conf:(1)>*
C. *data_channel_is_lifestyle=0 13863 ==> weekday_is_saturday=0 13863 <conf:(1)>*
D. *data_channel_is_lifestyle=0 13863 ==> weekday_is_sunday=0 13863 <conf:(1)>*
E. *data_channel_is_lifestyle=0 13863 ==> is_weekend=0 13863 conf:(1)*
F. *data_channel_is_lifestyle=0 ==> weekday_is_saturday=0 weekday_is_sunday=0 13863 conf:(1)*
G. *data_channel_is_socmed=0 weekday_is_saturday=0 13760 <conf:(1)>*
H. *data_channel_is_socmed=0 13760 ==> weekday_is_sunday=0 13760 <conf:(1)>*
I. *data_channel_is_socmed=0 13760 ==> weekday_is_saturday=0 13760 <conf:(1)>*
J. *data_channel_is_socmed=13760 ==> weekday_is_sunday=0 13760 <conf:(1)>*

VI. CONCLUSION

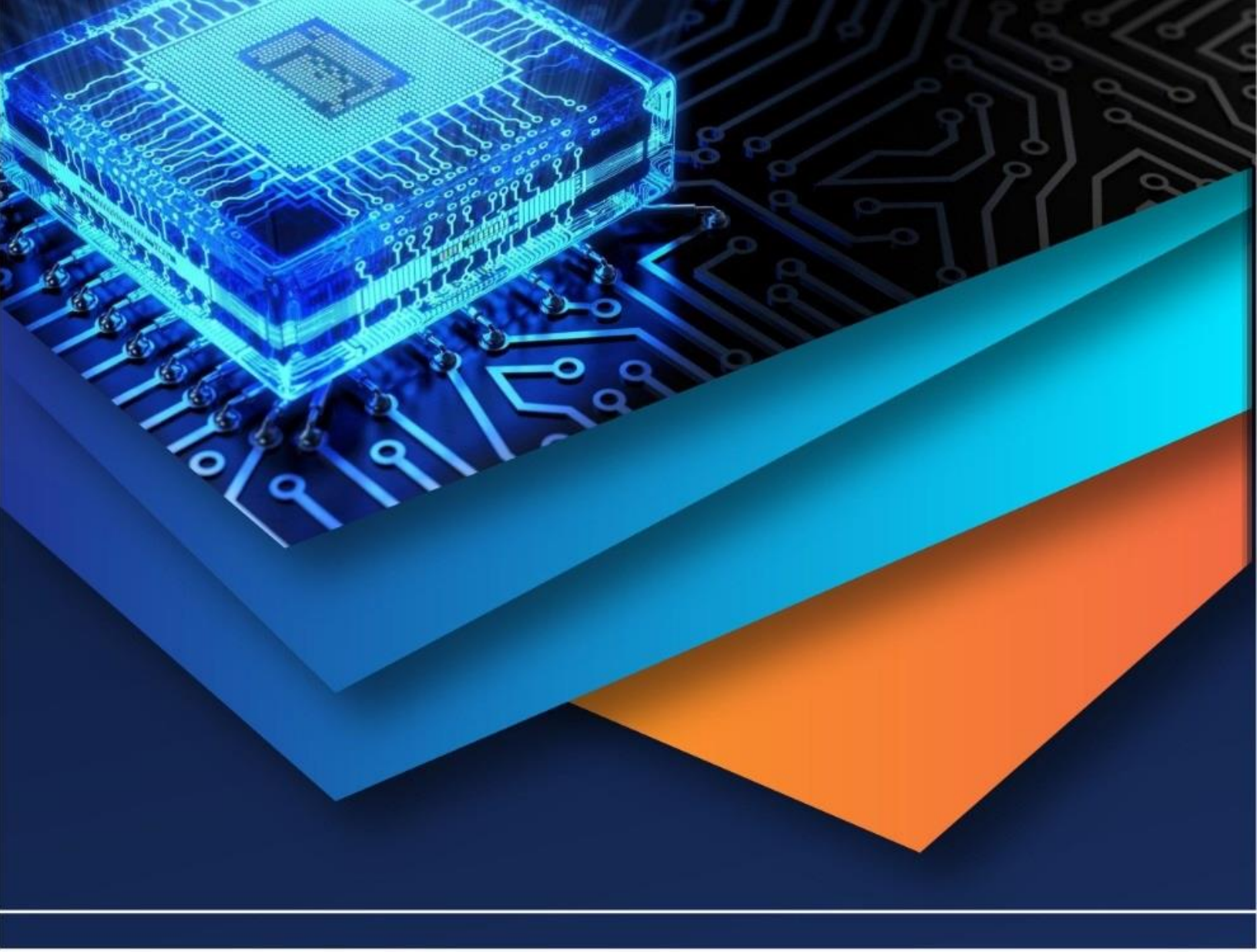
Test results shows that articles published on weekends especially of type ‘Lifestyle’ and ‘Social Media’ are usually the topic of interest for readers. From the results obtained, we have selected 10 best rules. This test result shows promise and is just a snapshot of some more interesting trends and patterns that we can discover by incorporating other factors (attributes) into the picture.
Such patterns can be used to rework the articles. Based on trends discovered, decision as to when an article should be published based on the category of the article, content subjectivity, polarity etc. This would also aid online advertisers and recommender systems in acting the way they should.

VII. SCOPE OF FUTURE WORK

The work proposed here can be improved by clubbing techniques like text mining and/or behavior analysis on the links along with social network analysis for creating an almost perfect solution for all the prediction needs.

REFERENCES

[1] Bandari R, Asur S, Huberman B (2012) The pulse of news in social media : forecasting popularity. Arxiv preprint arXIV: 12020332.
[2] Cosley, D.; Huttenlocher, D.; Kleinberg, J.; Lan, X.; and Suri, S. 2010. Sequential influence models in social networks. In 4th International Conference on Weblogs and Social Media.
[3] Jadhav R.; Kirutikha M. 2011. Pattern Discovery Using Association Rules. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011.
[4] Jamali S, Rangwala H (2009) Comment mining, popularity prediction, social network analysis. GMU-CS-TR-2009.
[5] Kempe, D.; Kleinberg, J. M.; and Tardos, ´ E. 2003. Maximizing the spread of influence through a social network. In KDD, 137–146. ACM.
[6] Lee, Moon, and Salamatian 2010), (Jamali and Rangwala 2009) and (Tatar et al. 2011) predict the popularity of a thread using features based on early measurements of user votes and comments.
[7] Tatar, A.; Leguay, J.; Antoniadis, P.; Limbourg, A.; de Amorim, M. D.; and Fdida, S. 2011. Predicting the popularity of online articles based on user comments. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11, 67:1–67:8. NewYork, NY, USA: ACM.
[8] Yu, B.; Chen, M.; and Kwok, L. 2011. Toward predicting popularity of social marketing messages. In SBP, volume 6589 of Lecture Notes in Computer Science, 317–324. Springer.
[9] Szaba'o., and Huberman, B.A. 2010. Predicting the Popularity of online content Commun. ACM 53(8):80-88.
[10] Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In ICWSM. The AAAI Press.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY