Popular Content Prediction Based on Web Visitor Data With Data Mining Approach

I D Iskandar¹,², N Ch Basjaruddin², D Supriadi³, Ratningsih⁴, D S Purnia⁵, and T Wibisono⁶

¹,₂,³,⁵,⁶ Universitas Bina Sarana Informatika. Indonesia
² Politeknik Negeri Bandung
³ Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
E-mail: iqbal.iql@bsi.ac.id

Abstract. A quality website has five parameters that must be considered are: information, security, convenience, comfort, quality of service. But of course, the fifth parameter does not always guarantee the amount through its Web page will increase, from that problem. So research is conducted to predict website content based on visitor data with a data mining approach, this research aims to improve the quality of content on target according to the interest of website visitors. Evaluation of random forest algorithm has the value the accuracy of classification of 71 percent by value of Kappa 0.712 whereas the k-NN algorithm has higher accuracy values of random Forest algorithm i.e. worth 84.88 percent and kappa values of the mean 0.847 the k-NN algorithm performs data processing process predictions against data of web content more effectively than the random forest.

1. Introduction
The website is an information system based on on-page, which is structured and connected. [1] on his research stated, “good quality website has five parameters to be concerned involved: information, security, practicality, comfort, and quality of service”. However, these matters don’t guarantee the number of web visitors. has concluded that for increasing the visitors’ interest, the website must update its content, clear interaction, trustworthy information, detail, and feedback service from the user. There are many well-designed websites to attract web visitors[2]. The different research method has been conducted can be applied to increase the traffic visitors on a web page to make a web page becomes well-identified popularity of online news from content meta data[3]. Besides, the good quality content of a website influence the visitors’ interest[4]. The five parameters of the website described are the condition of a good quality website, and the two of them are practical and comfortable. In relating to the statement[1], has researched the results of the researches conclude that correct link positioning influences the amount of click for accessing the web page, or its contents[5][6].

Similar research about predicting web page content that can be popular, prediction music track popular in online social networks[7]. Meanwhile researched to analyze how to popularize commercial websites[8] Solutions that have been gotten from those researches are classifying topics to be discussed, designing a good website, searching for content for the website from
products with famous brands. Nevertheless, the prediction pattern for deciding popular content in the future has not been available. Overall, the researches that have been studied only focus on the increase of website popularity by depending on website design, searching for causes that may influence the popularity of certain music, how users in online social network enjoy it and positioning link precisely to be clicked by visitors. Yet, they don’t discuss the increasing web visitors by implementing predicting method[9].

The reference of parameters on the research that had been conducted by Ren[7]is music acoustic content, music topic, artist reputation, and social context. Two titles by Ren’s literature studies, and Dimitrov’research, and some other sources that have been studied as well. The effort to boost popularity and quality of the web site still has minus, considering the researches are only focusing on website features quality and don’t focus on web content which should be relating to the theme. Once the relativity exists, the owners of the website will be easier to decide materials to be posted in favor of the visitors. Therefore, it is not only advertisements, which appear relating to visitors’ behavior but also material postings. Then, research will be conducted based on problems that occur regarding how to research predictions of popular web content based on website visitors with a data mining approach. the method used in this research is random forest and k-NN. With a decision tree, users can easily identify and see the relationship between the factors that influence a problem so that taking into account these factors can result in the best solution[10]. It’s simple and easily parallelized in other studies[11], the results of accuracy related to website content prediction using the random forest method indicate an accuracy value of 88.8 percent [12], while K-NN has an average accuracy value of 94 percent[13].

Then the indicative accuracy of the random forest and K-NN results from the experiment is used as a benchmark for use in this research. Considering the condition, there would be research-based on the problem involved, about how to do research predicting popular website content based on the click of its visitors, by implementing the Forecasting method to obtain problemsolving. This study aims to identify the decrease in website visitors every week and improve the decline by increasing website traffic with targeted quality content according to website visitors. This study research is going to test web etumbuhan.web.id since other websites may have a similar possibility for that case. The motivation and contribution of this research are to predict and identify the decline in website visitors every week on certain days and improve the decline by increasing website traffic with targeted quality content according to the interest of website visitors, besides this research can increase website traffic more visitors, and comparing which classification accuracy is best from two Decision tree and K-NN data mining algorithms.

2. Research Method

2.1. Random Forest

Random forest involves to Decision Tree implementing the bagging method. This model uses a hierarchy structure or tree structure [11]Once a decision tree is applied, users can identify and observe the correlation between factors to generate the best solution easily[14][15]. The decision tree can also analyze risk value and information value in problem-solving alternatives [12]. The following are steps of a random forest.

- Bootstrap step: extract sample randomly by using data training
- random sub-setting steps: arrange the tree based on the data but in each separating process, choose randomly m¡d distinction variable, and do the best separation.
- repeat the steps a-b as many ask to obtain k result of the random tree.
2.2. *k*-Nearest Neighbor (*k*-NN)

*k*-NN is a nonparametric method for classifying unknown data samples such as distance (by conducting Euclidean’s distance measurement) from the sample for another training example. The small ‘*k*’ is identified as distance and representative class in this class regarded as class output label. Commonly, ‘*K*’ value is decided by applying set -validation, or cross-validation[16]. *k*-Nearest Neighbor Rule works by searching for the amount of ‘*k*’ data object or the nearest training pattern with input pattern. Next, choose a class with a pattern as much as ‘*k*’. In short, *k*-NN results patterns by voting, and the nearest pattern is done based on the amount of ‘*k*’[17]. The following are *k*-NN algorithm steps:

- Add pattern x, f(x) into every training pattern.
- Add ‘*k*’ pattern with the nearest distance with X2 for input pattern X1 e.g. X1, X2 . . . , Xn
- Then, turn back the class with the highest amount of pattern among the ‘*k*’ pattern as decision class.

‘*k*’ parameter is the only parameter in k-NN functioning to direct generalization level toward the future data (out of training data). With a low parameter, k-NN will have high-level data generalization; conversely, generalization level will be low if the ‘*k*’ parameter is high and may classify data training only but failed in classifying the future data. Therefore, the parameter should be set correctly in the training process [18]. The structure of the predicting method of website content popularity will be proposed in Figure 1. [19] Predicting method of website content popularity is adopted by numerous of perfection in parts of the research process. [20]. The Empirical Refinement method applied to this research is aimed to identify problem-solving and minus from other research conducted as can be seen in research entitled, Visual Positions of Links and Clicks on Wikipedia[6][21].

![Figure 1](image)

**Figure 1.** The proposed structure of research activity process of the Empirical Refinement Method

The structure of the prediction method towards popular content are explained as follow:

1. **Domain Selection** The selected domain name is e-tumbuhan.web.id that is focusing on a variety of plantations since this research is aimed to predict popular content that has consistency with its theme.
2. **Exploring data content after publication** Historical data of the posting material after publication have been taken from google data analytic period November 26 until December 27, 2017. The data contains readers or web visitor’s interactions such as Page title, Day of week name, session duration, entrances, page view, and unique page view.
3. **Preprocessing** The prediction process will be held after preprocessing data for deciding relevant variables thus it results in a better model.
4. Data Cleansing  It is a process for detecting and correcting (omitting) damaged or inaccurate data on the data set table.

5. Prediction Method  Deciding a relevant method and suitable with desirable output requirement to predict popular content. The data mining forecasting method used in this research is K-nearest Neighbor Random Forest.

6. Prediction Pattern  A pattern of information in the form of information as a result of popular content prediction which has obtained after data containing visitors’ behavior is processed by the recommended algorithm, k-NN, and Random Forest.

7. Evaluation  The evaluation used uses a performance vector that includes the following criteria are added for polynomial classification tasks, Accuracy, Kappa statistics. Evaluation is conducted to measure the accuracy of the algorithm and Random Forest effectiveness.

3. Result and Discussion

Table 1. Sample Data Content after Publication

| Page Title       | Day of Session | Page Views | Unique Page View | Entrances |
|------------------|----------------|------------|------------------|-----------|
| Healthy Lemon    | Tuesday        | 625        | 15               | 2         |
| Healthy lemon    | Sunday         | 0          | 3                | 1         |
| Healthy lemon    | Thursday       | 4          | 2                | 1         |
| Plants 2012      | Monday         | 46         | 3                | 1         |
| Plants 2016      | Wednesday      | 21         | 9                | 1         |
| Plants 2017      | Tuesday        | 98         | 3                | 1         |
| Airmata pengantin| Wednesday      | 0          | 14               | 5         |
| Air matapengantin| Saturday       | 294        | 12               | 4         |
| air matapengantin| Sunday         | 1          | 12               | 3         |
| air matapengantin| Thursday       | 2          | 12               | 4         |
| air matapengantin| Friday         | 20         | 9                | 3         |
| air matapengantin| Tuesday        | 400        | 9                | 3         |
| air matapengantin| Monday         | 0          | 6                | 2         |

The contents of table 1 are examples of visitor data set posts from the website Etumbuhan.web.id is taken from Google Analytics, which will be processed using a data mining approach to get predictions of popular and unpopular content.

3.1. Prediction Method

Website data obtained from google analytic cannot be processed directly by using the data mining approach since the supporting attributes needed as a forecasting parameter of the popular website must be decided first.

Figure 2 is a prediction process method using the random forest algorithm and k-NN. After deciding the research domain, data set exploration, and deciding the algorithm next is building prediction methods. Decision tree pattern of popular content prediction a decision tree from the output of the decision tree algorithm. to simplify the classification process, data will be presented through table 2.

Table 2 is the data of the decision tree algorithm classification results in Figure 2, which will be explained as follows. According to the from prediction pattern of website’s popular contents theme explained on table 2, the contents which have average entrance higher than 13.5 percent,
Figure 2. The popular content prediction design process with rapidminer

Table 2. Popular content classification

| Entrance Page View | Unique Page View | Session Duration | Page Title        |
|--------------------|------------------|------------------|-------------------|
| >13.500 > 67.500   | > 14.500         | > 2663.500       | Citrus Lemon      |
| 14.500 > 90        | 23               | -                | Buni              |
| 13.500 75          | 23               | -                |                   |
| <13.500 < 67.500   | < 14.500         | > 1729.500       | Serealia          |
| 12.500 42          | 13.500           | -                | Anonna/sarikaya   |
| 12.500 35          | 3.500            | -                |                   |
| -                 | -                | > 1401 = < 1729.500 |                   |
| 3.500              | 20.500           | 3.500            | Yellow bamboo     |
| 3.500              | 23.500           | 3.500            | Burahol           |

page view 67 percent, unique page view higher than 4.5 percent from session duration more than 2663.500 seconds are available in content theme Citrus lemon and buni. After finding content data with popular topics, it then identifies unpopular content data by week, the results of which are presented in table 3.

Table 3 is a Classification based on Pageviews of less than 7 impressions every week. The classification results of the decision tree algorithm, based on data from the classification results, have found an unpopular website content topic and have several readers, from Monday to Sunday. The results of the classification can be concluded whether the content theme will continue to be made or not so that the posters can also avoid taking the theme of the content that has few publishers.

After knowing the classification of topics on websites that are not popular. Then the next step is to classify websites that have popular topics, based on the results of the decision tree algorithm data processing. so unpopular content can be replaced by creating content topics that will be popular from the results of the classification. Table 4 is the output for predicting website content that is potentially popular on the website Etumbuhan.web.id, to increase the number of page views article.

3.2. Evaluation and Validation of Algorithm

Data result of validation and evaluation of algorithm k-NN and Random Forest, explained in table 5 which states algorithm random forest has classification accuracy score toward variable for 71 percent with Kappa score, 0.712. Meanwhile, algorithm k-NN has a higher accuracy
Table 3. Classification based on Page Views less than 7 impressions every week

| Day of week name | Page Title         | Average page view |
|------------------|--------------------|-------------------|
| Monday           | Kokam              | > 4.500           |
| -                | Keledang           | < 4.500           |
| Tuesday          | Bunga Harum Sari   | > 4.500           |
| -                | Alamanda           | < 4.500”          |
| Wednesday        | Alokasia           | > 4.500           |
| -                | Aprikot            | < 4.500           |
| Thursday         | Jagung             | > 4.500           |
| -                | Melon Tanduk       | < 4.500           |
| Friday           | Alamanda           | > 4.500           |
| -                | Nanas              | < 4.500           |
| Saturday         | Pala- sakura cherry| > 4.500           |
| Sunday           | Pining Bawang      | > 4.500           |
| -                | Kastur             | < 4.500           |

Table 4. Classification based on Page View more than 60 impressions by week

| Nomor | Day of week name | Page Title                                                                 |
|-------|------------------|-----------------------------------------------------------------------------|
| 1     | Monday           | Plum, Citruslemon, Badam, Buni, Menteng, Blueberry                         |
| 2     | Tuesday          | CitrusLemon, menteng, durian, BungaBokor, sukun, alokasia, Tanamanpangan, Jambubol |
| 3     | Wednesday        | CitrusLemon, BungaBokor, Alpukat                                             |
| 4     | Thursday         | Badam, Buni, Kersen, bendu, Bunga Matahari, Jagung, keledang                |
| 5     | Friday           | Buni, Alokasia, CitrusLemon, Matoa, Blackberry, Alamanda, Buah              |
| 6     | Saturday         | Buni, Wresah, Badam, MelonTanduk, BungaBokor, Air Mata Pengantin, Menteng, Pala, Sukun |
| 7     | Sunday           | Badam, BungaBokor, healthylemon, Kepel                                      |

score than the Random Forest algorithm that is 84.88 percent with 0.847 Kappa score which means good in the prediction process. In conclusion, the respective data show that k-NN is more effective in the tabulation process of data prediction towards data content variables than Random Forest.

Table 5. Data result of Algorithm Evaluation

| Algorithm     | Accuracy | Kappa  |
|---------------|----------|--------|
| k-NN          | 84.88    | 0.847  |
| Random Forest | 71.60    | 0.712  |
4. Conclusion

The popular post prediction research is based on visitors on the website Etumbu.web.id with a data mining approach, the algorithm used for the prediction process is random forest and k-NN. The results of the research concluded that content themes that have an average entrance value of more than 13.5 percent, page views 67.5 percent, unique page views more than 4.5 percent, and session duration of more than 2663,500 seconds are found in the Citrus lemon and buni content themes, based on these data it means both of these content themes are more often accessed by web visitors. In addition, based on classification data from the k-NN algorithm and random forest, popular and unpopular website topics are known each week. so website owners can create content based on popular topics that are already known to correct unpopular topics. Suggestions for future research can be done by adding city and country variables so that popular content can be identified based on the classification of countries and cities that most access to the theme of the content contained on the website. Therefore the minus can be fixed. For future research, it is suggested to add the city variable and country variable. Thus, the number of cities and countries accessing the website’s popular content can be classified and noticed.

References

[1] Kim H and Niehm L S, 2009 The Impact of Website Quality on Information Quality , Value , and Loyalty Intentions in Apparel Retailing INTMAR 23 221–33
[2] Kim S, Kim S and Cho H 2011 Predicting the Virtual Temperature of Web-Blog Articles as a Measurement Tool for Online Popularity 449–54
[3] Uddin T, Jamshed M, Patwary A, Ahsan T and Alam, 2017 M S Predicting the Popularity of Online News from Content Metadata
[4] Yu J, Rui Y, Tao D and Member S 2019 Click Prediction for Web Image Reranking Using Multimodal Sparse Coding 23–32
[5] Dimitrov, 2017, What Makes a Link Successful on Wikipedia
[6] Dimitrov, 2016, Visual Position of Links and Clicks on Wikipedia
[7] Ren J, 2013, What Makes a Music Track Popular in Online Social Networks 95-6
[8] Dhokia, 2016 how to popularize commercial websites
[9] Xin J and Chen S, 2016 Bus Dwell Time Prediction Based on KNN 137 283–8
[10] Iskandar I D, Amirulloh I, Pertwi M W, Kusmiro M, Hikmah A B, Supriadi D, Bina U, Informatika S, Campus T, Author C and Tree D, 2020 Analysis Of Bubble Sort And Insertion Sort Algorithm On Memory Efficiency Using Data Mining Approach 16 89–96
[11] Iskandar I D, 2019 Parents’ Sum of Salaries Analyses towards School Tuition Fee Arrears Potential with Decision Tree Method Indones. J. Inf. Syst. 2 45
[12] Breiman L E O, 2001 Random Forests 5–32
[13] Shreyas R, Akshata D M, Mahanand B S, Shagun B and Abhishek C M, 2016 Predicting Popularity of Online Articles using Random Forest Regression
[14] Kamel S O M and Nour M, 2018 Feature Selection Methods For Predicting The Popularity Of Online news comparative study 6969–80
[15] Putri Z, 2019 Modified balanced random forest for improving imbalanced data prediction 5 58–65
[16] Lakshminarayanan B and Roy D M, 2013 Mondrian Forests Efficient Online Random Forests Proc. Int. Conf. Mach. Learn. (ICML), 1–9
[17] Peterson L E, 2019 K-nearest neighbor Characteristics of kNN 4 1–12
[18] Raikwal J S, Singhai R and Ph D, 2013 Integrating Markov Model with KNN Classification for Web Page Prediction 61 11–5
[19] Adeniyi D A, Wei Z and Yongquan Y, 2016 Automated web usage data mining and recommendation system using K-Nearest Neighbor ( KNN ) classification method Appl. Comput. Informatics 12 90–108
[20] Keshlholo Y, Wang S and Han E S, 2017 Predicting the Popularity of News Articles
[21] Deshpande D, 2017 Prediction and Evaluation of Online News Popularity using Machine Intelligence 2017 Int. Conf. Comput. Commun. Control Autom. 1–6
[22] Bunker R P and Thabtah F, 2017 Applied Computing and Informatics A machine learning framework for sport result prediction Appl. Comput. Informatics