Applying C-FDT as Making Decision for the Content of SEO Media Online

Erlin Windia Ambarsari¹,², Yulianingsih¹, Rendi Prasetya¹, Wahyu Nur Cholifah¹ and Robbi Rahim²

¹Departement of Informatics, Universitas Indraprasta PGRI, Jakarta, Indonesia
²Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia
³Media Have Fun, Banten Province, Indonesia

*erlin@mediahavefun.com

Abstract. The SEO is a tool for increasing traffic website through search engine result which has benefited, such as cost-effective and gives ROI (Return on Investment) based on many visitors. It is advanced research on content in Media Have Fun by using C-FDT on popular and trending articles found on Google Analytics and Yoast SEO. C-FDT is combined Fuzzy C-Means and J48. The reason is the content Media Have Fun has several attributes and data pattern which complex. The results obtained in this study are to improve SEO is the meta description, unique focus keywords, hashtags especially in the image. Therefore, visitors can easily do a page search.

1. Introduction

In Entertainment Media, the content is essential as the medium to spread information for readers, such as speeches, conferences, stage performances, and exhibitions. There are as known as MICE (Meeting, Incentive, Convention, and Exhibition) and it is usually used in the tourism industry activity [1], [2]. The role of Media itself to introduce to the public what interest the event is. Therefore, the challenges faced by journalists is to present information as attractive as possible on media content which the news still classified as hard news. The hard news itself depends on the Headline on the layout.

The first steps what journalist do are to find topics as news material. They do hunting by monitoring, interviewing, or taking pictures as article material — the article that was written by the journalist which is then submitted to the editorial for review before the news will publish. The editor is responsible for choosing which materials are considered attractive and editing articles that have been sent by journalists. The report is then set up to achieve SEO. SEO is the search engine optimization tool. Therefore, content can be performed on top of the searching, making it easier for readers to search for their desired topic.

In IT business, SEO has an advantage, one of which is as cost-effective and give long-term Return on Investment (ROI), although it has a long-term process which needs patience, time, money and correct implementation of SEO tools and techniques[3]. It requires a strategy to make a top content. Therefore, it needs to find out the behind reason why the content increase to the high, base on SEO.

As the previous study[4], to choose exciting content done by classifying it to determine the essential attributes, base on Google Analytics. The problem on the survey is out of sync with the data. Therefore,
this study will continue to try on add data of time series. Moreover, the data is not only from Google Analytics; also it utilizes Yoast SEO (WordPress SEO plugin). The purpose is to analyze traffic and the characteristic of content; the target is to compare popular and trending news contents.

The technique to analyze contents for traffic done by clustering attributes, such as Fuzzy Logic. The Fuzzy Logic is one of Artificial Intelligence method which has ranged between two options such as Yes and No, True and False, Good and Bad, and many more. The value to be between 0 and 1, the amount depends on the weight of the membership[5]. There is also a Fuzzy C-Means which functions to category more than one cluster which is the concept based on fuzzy c-partition[6]. Later on, to find out the characteristic of the media content, it needs to recognize the pattern and the feature for SEO Media Online before to making a decision. Several studies may combine with Fuzzy, such as Adaptive Neuro-Fuzzy Interface System (ANFIS)[7], [8] and Fuzzy Decision Tree (FDT)[9], [10].

Therefore, the content Media itself gained from Media Have Fun website as a sample has several attributes and data pattern which need investigate, then its analyze with using FDT, the Fuzzy method in this study will use is Fuzzy C-Means, that is C-FDT. The combine between Fuzzy C-Means and Decision Tree. The decision itself using the C4.5 algorithm, in Weka known as J48. J48 is an extension of the ID3 algorithm, its handle missing value and pruning trees[4].

2. Methodology Research
This work almost the same as previous research[4], which the step is do clustering the same type attributes and classification it with the decision tree. The method used as follows:

2.1. Fuzzy C-Means
This method works by making dataset (x_i) divided into several clusters, such as High, Moderate, and Low. Each data had given by random weight as the value of dataset member’s degree (\mu_i), which when it overall will be 1. Therefore, the equation [11] used is:
\[
\sum_{j=1}^{k} \mu_{ij} = 1
\]  
(1)

Which a cluster (c_j) formulated [11] as follows:
\[
0 < \sum_{j=1}^{k} \mu_{ij} < n
\]  
(2)

The value of membership’s degree for data x_i on cluster c_j, it is formulated by an equation [11] :
\[
\mu_{ij} = \frac{d(x_i,c_j)^w}{\sum_{i=1}^{k} d(x_i,c_j)^w}
\]  
(3)

Mentions: c_i is centroid cluster -jth, D(x_i,c_i) is range between data and centroid, and w is weighting exponent (which the value of w > 1).

Membership value will be inserted to the fuzzy pseudo-partition matrix, which the row is data and the column is membership value for each cluster.
\[
U = \begin{bmatrix}
\mu_{11}[x_1] & \mu_{12}[x_1] & \mu_{1k}[x_1] \\
\mu_{21}[x_2] & \mu_{22}[x_2] & \mu_{2k}[x_2] \\
\mu_{n1}[x_n] & \mu_{n2}[x_n] & \mu_{nk}[x_n]
\end{bmatrix}
\]  
(4)

While to calculate centroid on cluster c_i for feature j, it has equation [11] as follows:
\[
c_{ij} = \frac{\sum_{i=1}^{N} \mu_{ij}^w x_{ij}}{\sum_{i=1}^{N} \mu_{ij}^w}
\]  
(5)

Which N is overall data, w is weighting exponent and \mu_i as membership degree value x_i for cluster c_i. While the equation of objective function [11] is :
\[
f = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{ij}^w D(x_i,c_i)^2
\]  
(6)

2.2. C4.5
C4.5 is a decision tree method for use to classification various data mining and lead decision-making process [12]. A decision tree itself using a decision node which each path forms a decision rule to help identify a strategy for reach a goal with calculating conditional probabilities. C4.5 as improve ID3, such
as a possible use continues data, handling missing or unknown values, using attributes with different weights, and pruning the tree after created [13].

2.2.1. J48. J48 is a classification of simple C4.5 which creates a binary tree[14]. J48 is an implementation of the C4.5 in Weka using java which the classification is carried out recursively until every single leaf becomes obvious by producing the identity rules of specific data issued. The purpose is progressively generalized from the decision tree to obtain a flexible and accurate balance[4]. The J48 algorithm as follows [14]:

INPUT:
D    //Training data
OUTPUT
T    //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and label;
For each arc do
D= Database created by applying splitting predicate to D;
If stopping point reached for this path, then
T'= create leaf node and label with appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

According to [14], J48 to handling missing value by ignoring it, while to building a tree. The other words, it is based on known data for the attribute value in other records. The data divided into range based on the attribute values while the training.

2.3. C-FDT (Fuzzy C-Means Decision Tree)
According to subsection 2.1 and subsection 2.2.1, which they will be combined become C-FDT. The work’s step as follows:
a. Clustering set of viewer or visitor data per day (since September 1, 2018, until November 12, 2018) into cluster High, Moderate, and Low by Fuzzy C-Means.
b. Result Fuzzy C-Means combined to Yoast SEO data to classify data using the J48 decision tree.
c. For SEO data has considered by requirement as follow:
Table 1. SEO Requirement

| Symbol | SEO Categories |
|--------|----------------|
| A1     | A meta description has been specified, but it does not contain the focus keyword. |
| A2     | The focus keyword does appear in the SEO title. |
| A3     | Internal links appear in this page |
| A4     | The SEO title is longer than the viewable limit. |
| A5     | The images on this page are missing alt attributes. |
| A6     | No outbound links appear in this page, consider adding some as appropriate. |
| A7     | The focus keyword does not appear in the URL for this page. |
| A8     | The slug for this page is a bit long, consider shortening it. |
| A9     | The focus keyword appears in the first paragraph of the copy. |
| A10    | The exact-match keyword density, which is excellent; the focus keyword was found several times. |
| A11    | The meta description has a nice length. |
| A12    | The text contains a few words. The word is more than or equal to the recommended minimum of 300 words. |
| A13    | Never used this focus keyword before, |
| A14    | Not used the focus keyword in any subheading (such as an H2) in the copy. |

d. Noted that for SEO, its assessment of content is given value by using color as a fulfillment of requirements, the reason is for easy to understand. Including Gray means that the Admin has not chosen keywords, red is the ranking of keywords poorly, Orange or Yellow which means the keyword needs to be improved even though it does not affect the posts entirely, and Green means everything looks good.

3. Result and Discussion
The nine article content is taken as an example from Google Analytics in Media Have Fun which is influential and considered favorite are as follows:
a. /fakta-dari-cerita-greatest-showman
b. /film-jelita-sejuba-berkisah-tentang-seorang-istri-tentara-yang-hatinya-bergolak
c. /cti-it-infrastructure-summit-2018-hadirkan-teknologi-dan-best-practice-blockchain
d. /overwatch-season-8-sekarang-hadir-di-ps4-xbox-one-dan-pc
e. /pementasan-h2o-reborn-rupaka-sukses-digelar
f. /le-minerale-football-camp-siapkan-pesepak-bola-muda-indonesia-jadi-juara-nasional-internasional
g. /jiyuu-matsuri-2018-obake-yashiki-tampilkan-tempat-untuk-membunuh-diri-di-jepang
h. /tampil-maksimal-guns-n-roses-siap-guncang-jakarta
i. /kono-koi-wa-tsumi-nanoka-akan-hadirkan-drama

The results obtained from nine content where each content consists of 73 data sets with errors 2.85E-09. Where most visitors are $c_1$ (High), moderate visitors are $c_2$ and fewer visitors are $c_3$ (Low).
Table 2. Viewer Cluster

| µ   | c1     | c2     | c3        |
|-----|--------|--------|-----------|
| 1   | 1      | 3.11E-08 | 9.05E-09 |
| 2   | 2.56E-07 | 1      | 1.53E-07 |
| 3   | 0.001248 | 0.00263 | 0.996122 |
| 4   | 0.000974 | 0.001999 | 0.97028 |
| 5   | 0.001282 | 0.002628 | 0.996089 |
| 6   | 0.00841  | 0.016964 | 0.974626 |
| 7   | 0.001714 | 0.003487 | 0.994798 |
| 8   | 0.003993 | 0.008099 | 0.987908 |
| 9   | 0.000909 | 0.001852 | 0.997239 |

These results combined with SEO, where each attribute represented in Green, Yellow, and Red. Each color represented by the category on each attribute as follows:

Figure 1. Viewer SEO Attributes

The decision tree process found that for SEO results alone there is not much effect. However, some of the attributes that considered as increasing the popularity are as follows:

Table 3. Decision Tree

| Tree                           | Explanation                                                                 |
|--------------------------------|-----------------------------------------------------------------------------|
| A11 = Enoughlength: Short (3.0) | The meta description has a nice length. Length's maximals are twenty words and minimal four words |
| A11 = Nicelength: Long (6.0/1.0) | The focus keyword has unique which image on the page must appear alt attributes, Alt attributes are a hashtag on a picture |
| A13 = Unique                   | Add outbound links, for attracting visitors to view another page. However, this is not necessarily beneficial to content that gives the link. |
| A6 = Plenty: Fine (7.0/2.0)     | The focus is the keyword does appear in the URL in content. For the keyword itself, it should be unique |
| A6 = Adjust: Average (2.0)     |                                                                           |
| A7 = Notenough: Unique (3.0)    |                                                                           |
| A7 = Appear: Unique (4.0)      |                                                                           |
| A7 = Notappear: Common (2.0)   |                                                                           |

Based on the table above, it easily searchable content on search engines needs to determine the length of the meta description, the meta description function as a brief description of a page's content.
Therefore, a reviewer can understand a brief explanation of the contents of the information before they visit the page.

Also, the need for the right keywords (focus keywords), if the keyword is unique, it allows it to remain in the competition, based on the chosen topic. Image hashtags need to be built because clicking on a picture can make it easier for visitors to get the page URL.

For the implementation of C-FDT in this study, there has not been a significant effect on overall SEO, especially in its exact measurement (based on color to achieve SEO needs). However, for the main focus on solving the direction of SEO can be done.

4. Conclusion
According to C-FDT, what needs to be considered to improve SEO is the meta description, unique focus keywords, hashtags especially in the image. Therefore, visitors can easily do a page search. Although C-FDT cannot yet classify the whole of SEO, some essential categories can still be analyzed as decision making. Therefore, its need another measurement of specific, such as extended time visitors, inspiring articles based on the viewer.

References

[1] Smagina N, 2017 The internationalization of the Meetings-, Incentives-, Conventions- and Exhibitions- (MICE) industry: Its influences on the actors in the tourism business activity J. Econ. Manag. 27, 1 p. 96–113.

[2] Pearlman D M and Mollere L, Sep. 2009 Meetings, Incentive, Convention, and Exhibition Evaluation Practices: An Exploratory Study among Destination Marketing Organizations J. Conv. Event Tour. 10, 3 p. 147–165.

[3] Sohail A, 2012, Search Engine Optimization Methods & Search Engine Indexing for CMS Applications, Lappeenranta University of Technology.

[4] Windia Ambarsari E. 2018 Klasifikasi Daya Tarik Konten Artikel Media Daring Dari Data Google Analytics Dengan C-FDT J. Inform. J. Pengemb. IT 03, 02 p. 211–218.

[5] Indahingwati A Barid M Wajdi N Susilo D E Kurniasih N and Rahim R, 2018 Comparison Analysis of TOPSIS and Fuzzy Logic Methods On Fertilizer Selection Int. J. Eng. Technol. 7, 2.3 p. 109–114.

[6] Parwatiningtys D Ambarsari E W and Mariko S, 2017 The calculation of the highest leak level of water pipe lines region at PDAM Tirta Kahuripan using fuzzy C-means and ArcGIS method analysis AIP Conf. Proc. 030196.

[7] Dhika H Daengs GS A and Ambarsari E W, 2018 Forecasting Water Loss Due To Pipeline Leakage By Using ANFIS And BACKPROPAGATION Approach (Study Case At PDAM Tirta Kahuripan On District of Bogor) in Proceedings of the Joint Workshop KO2PI and The 1st International Conference on Advance & Scientific Innovation p. 119–125.

[8] Afriliana I Budihartono E and Maulana A, Jul. 2017 Pengukuran Kinerja Dosen Menggunakan Metode Adaptive Neuro Fuzzy Inference System (ANFIS) J. Inform. J. Pengemb. IT 2, 2 p. 109–112.

[9] Segatori A Marcelloni F and Pedrycz W, Feb. 2018 On Distributed Fuzzy Decision Trees for Big Data IEEE Trans. Fuzzy Syst. 26, 1 p. 174–192.

[10] Narayanan S J Paramasivam I and Bhatt R B, 2017 On the estimation of optimal number of clusters for the induction of fuzzy decision trees Int. J. Data Sci. 2, 3 p. 221.

[11] Astria D and Suprayogi S, Mar. 2017 Penerapan Algoritma Fuzzy C-Means Untuk Clustering Pelanggan Pada CV. Mataram Jaya Bawen Eksplora Inform. 6, 2 p. 169–178.

[12] Dai W and Ji W, 2014 A mapreduce implementation of C4.5 decision tree algorithm Int. J. Database Theory Appl. 7, 1 p. 49–60.

[13] HSSINA B MERBOUHA A EZZIKOURI H and ERRITALI M, 2014 A comparative study of
decision tree ID3 and C4.5 *Int. J. Adv. Comput. Sci. Appl.* **4**, 2.

[14] Patil T R and Sherekar, S S, 2013 Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification *Int. J. Comput. Sci. Appl. ISSN 0974-1011* **6**, 2 p. 256–261.