Journal of Economics, Technology and Business (JETBIS)
Volume 3, Number 11 November 2024
p-ISSN 2964-903X; e-ISSN 2962-9330

Utilization of Query Expansion Using Data Mining Method In Analyzing Documents on The Irama Nusantara Website

Rizky Aulia[1], Agung Mulyo Widodo[2]
Universitas Esa Unggul, Indonesia
*e-mail: henry.arianto8@gmail.com
*Correspondence: henry.arianto8@gmail.com

KEYWORDS:
Machine Learning, Natural Language Processing, Big Data, Data Mining

ABSTRACT

In Indonesia, many local websites, such as Irama Nusantara, hold valuable information related to music and culture. Although rich in data, the utilization of this information is still limited. This research aims to utilize query expansion techniques through data mining methods in analyzing data from the Irama Nusantara website. Data was collected from the Irama Nusantara website through a crawling process, resulting in 5404 entries covering audio, images and text. The analysis was conducted using Natural Language Processing (NLP) techniques starting with the preprocessing stage. Next, the K-Means algorithm was applied for clustering, and the Term Frequency-Inverse Document Frequency (TF-IDF) method was used for term weighting. Classification models were built using Support Vector Machine (SVM) and Naive Bayes for comparison. The analysis shows that the use of query expansion significantly improves the accuracy of information retrieval on the Irama Nusantara website. The method evaluation showed that SVM gave better results in terms of accuracy and precision compared to Naive Bayes. In addition, Principal Component Analysis (PCA) shows that 70-95% of the variance in the data can be explained by the resulting principal components, which signifies the efficiency of the applied method. This research not only provides a deeper insight into the patterns and trends in the analyzed data, but also contributes to the development of information technology in the field of culture in Indonesia. This research successfully developed an effective analysis model to utilize data from the Irama Nusantara website.

INTRODUCTION

In today's digital age, the amount of data generated globally is increasing rapidly, creating huge challenges in terms of management and analysis (Chen & Zhang, 2014; Leeflang et al., 2014; Moorthy et al., 2015; Sivarajah et al., 2017). According to an IDC report, in 2020, the global data volume has reached 44 zettabytes and is expected to grow to 175 zettabytes by 2025 (Bar-Lev et al., 2024; Cha, 2023; Ghosh et al., 2023). This massive availability of data creates an urgent need for effective analysis techniques and methods, including the use of data mining and machine learning to transform data into useful information.

In Indonesia, the challenge of managing big data is also very much present, especially in the context of utilizing information available on local websites (Abdillah et al., 2024; Carley et al., 2016; Harakan et al., 2024; Hilbert, 2016). The Irama Nusantara website, which provides a wide range of information related to music and culture, is one such rich yet underutilized data source.

The lack of effective methods to analyze and extract information from this website has resulted in the potential value of the data not being used optimally.

Several previous studies have examined the application of data mining in different contexts, but few have focused on utilizing data from local websites such as Irama Nusantara. Previous research has shown the importance of query expansion in improving the accuracy of information retrieval, but has not applied the method to the context of music and culture in Indonesia (Goyal et al., 2018; Jena & Rautaray, 2019; Kambau & Hasibuan, 2017; Nie, 2022). Thus, there is a research gap that needs to be filled to explore the potential of data on these sites.

Given the importance of the information that can be obtained from the Irama Nusantara website, this research is urgent to do. By utilizing the existing data, it is expected to increase the accessibility of information for users and provide deeper insights into music and cultural trends in Indonesia (Lee et al., 2020; Saragih, 2023; Spiller & Clendinning, 2022). In addition, utilizing data with the right methods can support better decision-making in the music industry.

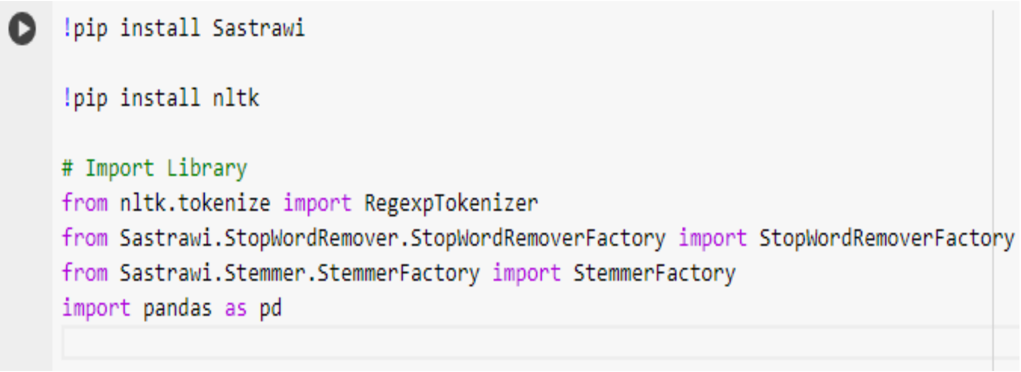
This research offers a new approach by applying query expansion methods using data mining and machine learning techniques to analyze data from the Irama Nusantara website. By integrating Natural Language Processing (NLP) and clustering algorithms such as K-Means, this research seeks to produce a more in-depth and relevant analysis. This approach differs from previous research which is more general in nature and does not focus on the local context.

The main objective of this research is to develop an effective analysis model in utilizing data from the Irama Nusantara website. This research aims to crawl the data, perform preprocessing, and apply machine learning techniques to improve the relevance and accuracy of the information obtained. Thus, it is expected to make a real contribution to the development of information technology and culture in Indonesia.

The expected benefits of this research include an increased understanding of the use of big data in a cultural context and the provision of tools that can be used by researchers and practitioners in the music industry. In addition, the results of this research can be used as a reference for future information system developers to improve the quality of services and accessibility of information for the wider community. Thus, this research not only contributes to academic development but also provides a significant social impact.

RESEARCH METHOD

Data collected online from the Irama Nusantara website in the form of secondary audio, image, and text data as many as 5404 entries are used to find the initial query from the user session, and then recreate a series of actions in the session. Data analysis method using Machine Learning Model, Text Preprocessing. The next stage is to do preprocessing. The result of this preprocessing is an index of all terms on the Irama Nusantara website. Previously, the author will import the library module from Sastrawi with the following coding.



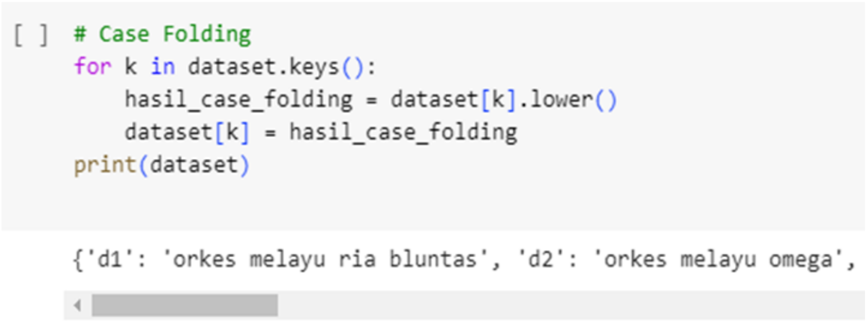
```
!pip install Sastrawi

!pip install nltk

# Import Library
from nltk.tokenize import RegexpTokenizer
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import pandas as pd
```

In the case folding stage, the letters are converted to all lowercase with the aim of facilitating text identification in each preprocessing. In the case folding process, the computer reads lowercase and capital letters in a text as different letters, in order to avoid case sensitivity, of course the preprocessing results will not be optimal, the purpose of this process is to facilitate the next process.

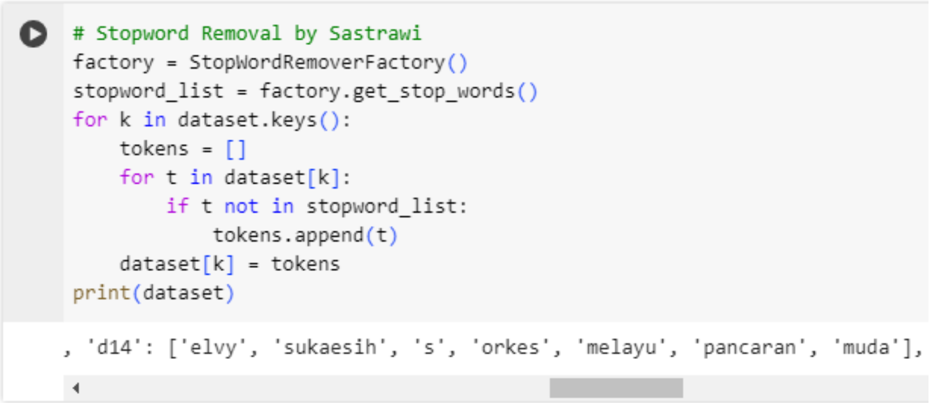
Before *case folding*	Top Hits Lagu-Lagu Panbers
After *case folding*	top hits lagu-lagu panbers



```
[ ]  # Case Folding
     for k in dataset.keys():
         hasil_case_folding = dataset[k].lower()
         dataset[k] = hasil_case_folding
     print(dataset)

     {'d1': 'orkes melayu ria bluntas', 'd2': 'orkes melayu omega',
```

Filtering is done to remove punctuation, double spaces, new lines and stop words because they do not have an intention. In addition, words that do not have an intention can be conjunctions (if, and, when) or prepositions (in, to, on).



```
# Stopword Removal by Sastrawi
factory = StopWordRemoverFactory()
stopword_list = factory.get_stop_words()
for k in dataset.keys():
    tokens = []
    for t in dataset[k]:
        if t not in stopword_list:
            tokens.append(t)
    dataset[k] = tokens
print(dataset)

, 'd14': ['elvy', 'sukaesih', 's', 'orkes', 'melayu', 'pancaran', 'muda'],
```

The query entered as input can be expanded with query expansion. The first step to expand the query is to rearrange the query and add words that have the same meaning (synonyms) and then pair them with the initial query. The list of synonyms to be taken comes from the Indonesian thesaurus dictionary. The query expansion process can be seen in the following flowchart. Synonyms are taken up to three words from the Indonesian thesaurus dictionary for a specific word. An example of a synonym dictionary that was built can be seen in the following table.

Word	Synonym
Voluntary	Relawan, Sukarelawan, Volunteer, Susul
Instant	Brief, Brief, Brief, Brief
Sing	Humming, singing, singing,
Miss	Love, Mind, Sorrow

Source: Processed Data

After preprocessing, the data will produce a set of terms that are neatly arranged. Each term will be counted the number of times it appears in the data. The calculation result of each term is called the term frequency and will be stored in the database for the next process, namely the calculation of the probability of each term that matches the query. Then, the indexing process is carried out, namely storing a number of terms in the document so that the system does not need to run preprocessing every time the user enters a query.

RESULTS AND DISCUSSION

This section will explain the implementation of each step that has been explained in the previous chapter and present the results of the trials that have been carried out according to the test scenario, namely by comparing the level of accuracy and precision between the methods used in this study with the methods used in previous studies. Then at the end of the chapter will be presented the evaluation and discussion of the experimental results obtained. The input data from this study is in the form of document queries on the Irama Nusantara website, then the preprocessing process, weighting, semantic calculations and similarity calculations are carried out, and the output is in the form of ranking according to the query searched for by the user. In integrating the model, the method used is implemented in planned steps so that measurable test results are obtained. The programming language used is Python in the Google Colab application.

Table 1. ***Weighting Index TF -IDF***

Term	TF-IDF
Musica	0.303216
nur	0.42616
asiah	0.42616
djamil	0.42616
ba	0.42616
resah	0.42616

Source: Processed Data

Table 2. ***TF-IDF score in term total***

Term	TF-IDF
musica	0.278943
tom	0.392044
slepe	0.392044
amit	0.784088
amit	0.784088

Source: Processed Data

Table 3. ***Score Euclidean d1-d2***

Score Euclidean Distance d1	0	1.353085
Score Euclidean Distance d2	1.353085	0

Source : Processed Data

Table 4. ***Cosine Similarity Score d1-d2***

Score Cosine Similarity d1	1	0.08458
Score Cosine Similarity d2	0.08458	1

Source : Processed Data

CONCLUSION

If the cosine similarity value is close to 0, it indicates that the documents are not similar. This can happen when the angle between the TF-IDF vectors is greater than 90 degrees. In this experiment, the combination of inference and learning has found useful groups in the dataset owned by Irama Nusantara. Overall, interpreting the PCA score with the variance ratio that has been given by the system represents the proportion of variance explained by each principal component obtained from the PCA (Principal Component Analysis) procedure. The cumulative variance ratio that can be explained is 70-95%. Naive Bayes classification is determined by the density function of the probability that is converted into a posteriori or the probability that follows its class/measurement, namely, the probability after measurement. SVM represents training examples as points in dimensional space, then mapped so that examples of data classes are separated by a dimensional hyperplane chosen to maximize the "margin" on both sides of the hyperplane. A high precision value will explain that k-fold cross validation achieves the desired value. k fold cross validation accuracy will calculate how close the measured values are to the true values so that it will be considered accurate, and can generate a set of features, and run the learning algorithm using only these features, and evaluate the resulting classifier using k fold cross validation (or single holdout set).

BIBLIOGRAPHY

Abdillah, A., Widianingsih, I., Buchari, R. A., & Nurasa, H. (2024). Big data security & individual (psychological) resilience: A review of social media risks and lessons learned from Indonesia. *Array*, 100336.

Bar-Lev, D., Sabary, O., & Yaakobi, E. (2024). The zettabyte era is in our DNA. *Nature Computational Science*, 1–5.

Carley, K. M., Malik, M., Landwehr, P. M., Pfeffer, J., & Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. *Safety Science*, *90*, 48–61.

Cha, J. (2023). Big Data Studies: The Humanities in Uncharted Waters. *Korean Studies*, *47*(1), 274–299.

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347.

Ghosh, P., Biswas, A., & Ghosh, S. (2023). Fundamentals and Technicalities of Big Data and Analytics. In *Intelligent Systems in Healthcare and Disease Identification using Data Science* (pp. 51–106). Chapman and Hall/CRC.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21–43.

Harakan, A., Abdillah, A., Said, T. G., Mujizatullah, M., & Gray, S. (2024). Big Data and Security: A Review of Social Media Risks and Insights for Indonesia. *Journal of Governance and Public Policy*, *11*(1), 14–32.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1), 135–174.

Jena, G., & Rautaray, S. (2019). A comprehensive survey on cross-language information retrieval system. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, *14*, 127–134.

Kambau, R. A., & Hasibuan, Z. A. (2017). Unified concept-based multimedia information retrieval technique. *2017 4th International Conference on Electrical Engineering, Computer*

Science and Informatics (EECSI), 1–8.

Lee, Y. L., Jung, M., Nathan, R. J., & Chung, J.-E. (2020). Cross-national study on the perception of the Korean wave and cultural hybridity in Indonesia and Malaysia using discourse on social media. *Sustainability*, *12*(15), 6072.

Leeflang, P. S. H., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12.

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big data: Prospects and challenges. *Vikalpa*, *40*(1), 74–96.

Nie, J.-Y. (2022). *Cross-language information retrieval*. Springer Nature.

Saragih, H. S. (2023). Predicting song popularity based on Spotify's audio features: insights from the Indonesian streaming users. *Journal of Management Analytics*, *10*(4), 693–709.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286.

Spiller, H., & Clendinning, E. A. (2022). *Focus: Gamelan Music of Indonesia*. routledge.