An Approach to Recommend Fishing Location and Forecast Fish Production by Using Big Data Analysis and Distributed Deep Learning

Minh-Triet Thai*,†, Thao-Ngan Chu-Ha*,†, Tuan-Anh Vo*,†, Trong-Hop Do*,†

* Faculty of Information Science and Engineering, University of Information Technology
† Vietnam National University, Ho Chi Minh City, Vietnam.

***Abstract*—In fishing industry, fish populations can move over a vast sea and thus boats often search for days or weeks before making a catch. Given the excessive CO2 emissions from vessels and rising fuel cost, it is important to optimize commercial fishing activities by reducing unnecessary searching period. This study proposes a hybrid recommendation system for predicting the best locations for catching specific fish species and experimented various deep learning based multi-variate time series models to forecast the gross weight of two important species of Norwegian fishery: haddock and mackerel. To ensure the practicality, both recommendation system models and time series forecasting models are trained and deployed using Apache Spark and BigDL, which are frameworks for big data processing and distributed deep learning training. The proposed hybrid recommendation model achieve good performance with RMSE of 0.4933. Deep learning based models are also shown to achieve high performance on forecasting fish gross weight. Through this study, it is revealed that datetime and environmental features can play important roles for building sustainable commercial fishing plan.**

***Index Terms*—Fishing forecast, Big Data, Time Series, Recommendation System, Distributed Deep Learning**

I. Introduction

With a fishing zone spanning 2.1 million square meters, Norway is considered Europe's largest fishing and aquaculture nation. Every year, commercial vessels catch fish with a total value of around 20 billion NOK from the Norwegian fishing zone. Planning offshore fishing trips for research or commercial objectives always requires careful consideration of supplies, fuel, and especially precision for locating fish. The overall migration patterns of the major fish species in the area are relatively predictable and common knowledge. However, fish populations can move over a vast sea which makes boat to search for days or even weeks before making a catch. Beside the excessive CO2 emission from vessels and rising fuel cost, the searching process also introduce negative impacts on marine ecosystem. Therefore, it is ideal to be able to decide the optimal fishing location before the trip to avoid unnecessary searching. Such decision can be made based on the predictions of the optimal fishing locations for specific fish species and the output in several days in future.

The contribution of this paper is two-fold. First, we developed a hybrid recommendation system for suggesting the fishing locations a vessel should prioritize in order to maximize the likelihood of catching a specific fish species. Second, we applied deep learning-based forecasting models to predict the gross weight average of two commercially important species of the Norwegian fishery: haddock (Melanogrammus aeglefinus) and Norway mackerel (Scomber scombrus). The input data used for the recommendation and forecasting models are daily catch nodes provided by the Norwegian Fishing Directorate and environmental data (sea surface temperature).

Since datasets have been collected for many years, it might be too large to be used for training recommendation and forecasting models using traditional machine learning libraries. Also, the practicality of the system requires the trained models to process a huge amount of input data to produce results for various fish species on a vast sea. Therefore, to ensure the practicality of the system, all recommendation models and forecasting models are trained and deployed using big data technology. Specifically, recommendation models are trained using Spark MLlib. The deep learning based forecasting models are trained in distributed manner thanks to BigDL library. All models are deployed on Spark and thus the system can be scaled up to process any amount of data required by the fishing planing aplication.

II. Related Works

There are studies developed recommendation system for many topics that involve huge amount of data in daily basis, including movie recommendation on MovieLens dataset [1], Netflix Prize [2] or book recommendation [3]. Different techniques and methods has been used to build these recommendation systems, including content-based filtering [4], collaborative filtering [5], [6], hybrid recommender with weighted combination technique [7]. To our knowledge, none of these approaches have been applied to recommend the fishing locations given a type of fish.

Other approaches exist that forecast the spatio-temporal distribution of fish in relation to environmental conditions [8], [9] Many of these studies predict the probability of fish presence at a given location at sea. The methods usually consist of two main blocks: predicting the sea temperature through satelite image or time series data, then based on the predicted sea temperature in different periods, the migration position of the fish school is predicted accurately using difference type of machine learning models.

III. Dataset

A. Catch Notes Dataset

The core dataset used in this project is a collection of catch notes compiled by the Norwegian Fishing Directorate [1] from the years 2000 to May 2022. The notes consist of information about the catch that is manually logged during landing, e.g., when it was caught, where it was caught, what equipment was used, and the species distribution of the catch. Each observation contains 133 data fields and around one million notes each year, which forms a large dataset with 26,032,921 records in total.

Features of our interest include information about the locations of the catch, information about fish caught (species code, gross weight) and types of fishing gear used in the catch. Therefore, the dataset can be filtered and reduced in size depending on the properties of each experiment. The processes including feature selection and data preparation during experiments with this dataset are described in detail in the next section.

B. Sea Surface Temperature Data (SST)

An increase in temperature is likely to affect the timing and magnitude of the growth, recruitment and migration of North Sea species, such as haddock and Norwegian mackerel, with subsequent impacts on its sustainable exploitation. Consequently, it is critical to collect sea temperature data in order to gain a better understanding of fish migration, which will aid in accurately predicting species' gross weight.

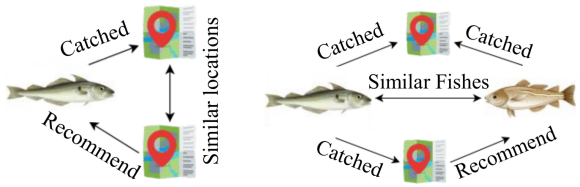
The temperature data used in the experiments is provided by National Oceanic and Atmospheric Administration (NOAA, US). It contains information about the daily estimate sea surface temperature in degree Celsius globally and stored under netCDF4 formats. The data was collected from satellite observations, and consists of daily data at 0.25 degree latitude × 0.25 degree longitude resolution. We only use the subset of data from year 2000 to 2022 in our experiments to fit with catch notes data.

[1]https://www.fiskeridir.no/Tall-og-analyse/AApne-data/Fangstdata-seddel-koblet-med-fartoeydata

IV. Proposed methods

A. Recommendation System

Two type of recommendation systems: Content-based Filtering and Collaborative Filtering, which are illustrated in Fig. 1, are used to get seperated set of locations. The two sets are combined to create the final set of recommended locations.



(a) Content-based filtering (b) Colaborative filtering

Fig. 1. Fishing locations recommendation systems

1) Content-based Filtering: Content-based filtering (CB) makes recommendations for fishing locations with similar content to those based on fish activity in the past. For example, based on the conditions of catching a type of fish, such as equipment used in referenced locations. To suggest that other areas with the same equipment could find the exact fish species and catch. Figure 1(a) shows an example of Fishing locations recommendation using content-based filtering.

The CB method requires grouping fishing locations with similar attributes, considering the fish species, and then seeking those terms in the dataset. Finally, we suggest different fishing locations with match content.

2) Collaborative Filtering: Collaborative Filtering (CF) is a method of analyzing an object's data to find the correlation between objects. In this problem, the collaborative filtering method aims to estimate the likelihood of catching specific fish in a location based on other fish with similar properties and behaviors. The determination of "similar" between fish species can be based on a rating of the likelihood of catching this fish at a location in the past.

As an example in Fig. 1(b), two fish species are frequently caught in the same location and have high rating. Considering the history of the system, a fish species is also caught at a new location, so the system recommends that new location could be a good fishing location for the another fish species.

3) Hybrid Recommendation: For recommender systems, collaborative filtering is the preferred method since it is completely independent of the representation of the objects being suggested, however, there are still restrictions called 'cold start' for new species. In other words, collaborative filtering requires large amounts of existing data to make accurate recommendations, while new species have limited data accuracy. In contrast, content-based filtering can understand species based

on their existing information. This requires a hybrid recommendation as a solution. In order to reduce the drawbacks of each approach, the hybrid recommender, as shown in Fig. 2, combines content-based filtering and collaborative filtering models, providing recommendations for the optimal location for each species.

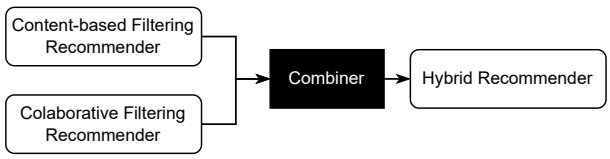


Fig. 2. Fishing locations recommendation using Hybrid Recommendation

We will present these approaches in detail in the Experiments Setup section.

B. Time Series Forecasting Models

1) Temporal Convolutional Network: Temporal Convolutional Network or simply TCN, is a variation of Convolutional Neural Networks for sequence modelling tasks, by combining aspects of RNN and CNN architectures. Preliminary empirical evaluations of TCNs have shown that a simple convolutional architecture outperforms canonical recurrent networks such as LSTMs across a diverse range of tasks and datasets while demonstrating longer effective memory.

The convolutions in the architecture are causal, meaning that there is no information "leakage" from future to past. The architecture can take a sequence of any length and map it to an output sequence of the same length, just as with an RNN. TCNs possess very long effective history sizes (i.e., the ability for the networks to look very far into the past to make a prediction) using a combination of very deep networks (augmented with residual layers) and dilated convolutions. Causal Convolutions enables large scale parallel computing which makes TCN has less inference time than RNN based model.

2) Sequence-to-Sequence Model: Seq2Seq model with Encoder-decoder architecture have provided state of the art results in sequence to sequence NLP tasks like language translation, etc. Multistep time-series forecasting can also be treated as a seq2seq task, for which the encoder-decoder model can be used.

The Seq2Seq model we used in this report is based on basic version of LSTM - VanillaLSTM - that has a single hidden layer of LSTM units, and an output layer used to make a prediction, it is suitable for multivariant and multistep time series forecasting.

V. Experiments

All of the experiments within two tasks: fishing location recommendation and gross weight forecasting are implemented in Python 3.7.13 within Google Colaboratory using PySpark - a Python interface for Apache Spark engine for dealing with big data and BigDL framework for developing time series models.

A. Fishing Locations Recommendation

1) Features Selection and Preprocessing: In order to develop fishing location recommendation systems, features from Catch notes dataset are used as materials for the experiments. The construction of recommendation systems requires a rating standard to describe the relation between locations and fish species. Since the dataset does not contain any specific discrete feature such as rating, we have to investigate creating a similar one from the available features.

The population of fish in an area can be reflected by the species gross weight of each catch. If the gross weight of a species is high in the catch made in a specific location, it indicates that the location provides ideal environmental conditions for the growth of the fish population and is considered a good fishing spot and reverse. Therefore, we determine to use the fish's gross weight from the catch to rate the species distribution in a certain location.

We first remove missing values from catch notes data and perform a discretization process to map continuous fish's gross weight into discreate rating values using our own standard. The mapping rule of discretization process and the distribution of ratings are shown in Table I and Fig. 3, respectively.

TABLE I
PROPOSED MAPPING STANDARD FROM GROSS WEIGHT TO RATING

Gross weight(kg)	Rating	Gross weight(kg)	Rating
< 10	0.5	1,000 – 5,000	3
10 – 50	1	5,000 – 10,000	3.5
50 – 100	1.5	10,000 – 50,000	4
100 – 500	2	50,000 – 100,000	4.5
500 – 1,000	2.5	$\geq 100,000$	5

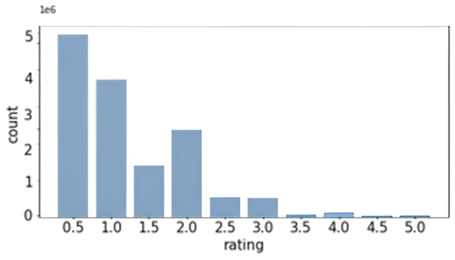


Fig. 3. Distribution of Ratings (Gross Weight)

After the discretization process, we choose four features from the Cactch notes dataset for developing recommendation systems, including:

- **Species FDIR code**: Unique ID for fish species in Norwegian Directorate of Fisheries' code list.

- **Location code**: Unique ID for fishing location.
- **Gear code**: Unique ID for equipment used in the catch.
- **Rating**: Rating of fish's gross weight of the catch.

Because there are many cactch notes contain the same information about the species, location and gear used in the catch, but different gross weight, we group by those features and calculate the average rating. The data preparation is complete and ready to be used in the experiments.

2) Experiments Setup:

a) Content-based Filtering: In this method, we recommend fishing locations for a type of fish based on the gears used in the catch. First, we collect set of all gears code and calculate the average rating once again for all gears used to catch specific fish in each location. The data after the process is described in Table II.

TABLE II
GROUPBY SPECIES, LOCATION AND PROCESS ON GEAR CODE AND RATING

Species_code	Location_code	Gear_code	Rating_avg
211	1	[51, 20, 61, 22, 55]	1.1603...
211	2	[20, 35, 32, 22]	1.1875...
211	3	[51, 20, 35, 32, 22, 55]	0.9791...
211	4	[33, 51, 20, 35, 21, 22, 55]	0.9829...
211	5	[51, 35, 22]	1.7256...

Then, we explode the Gear_code column, sort and transform it into one-hot vectors. The gear codes used to catch a specific fish in an area is presented by a vector with length of 40 and used for calculating the similiarity. The resulted dataframe is show in Table III. After this step, we randomly split the dataframe into train set and test set with ratio 8:2 for evaluation.

TABLE III
TRANSFORM SET OF GEAR CODE INTO ONE-HOT VECTOR

Species_code	Location_code	Gear_code	Rating_avg
211	1	[0, 0, 0, 0, 1, 0, 1, 0, ...]	1.1603...
211	2	[0, 0, 0, 0, 1, 0, 1, 0, ...]	1.1875...
211	3	[0, 0, 0, 0, 1, 0, 1, 0, ...]	0.9791...
211	4	[0, 0, 0, 0, 1, 1, 1, 0, ...]	0.9829...
211	5	[0, 0, 0, 0, 0, 0, 1, 0, ...]	1.7256...

For each species and a given location in test set, we proceed to compare the gears used in the location with gears used at all locations catched that species in train set. Consine similarity formular, as shown in Equation 1, is used to calculate the distance between gear code vectors. Thereby, we can explore which locations in train set use the most similar equipments with the given location in test set to make the predictions. Finally, the predicted rating for a location to catch a specific fish in test set is calculated by the average ratings of similar locations catched that fish in train set.

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}} \quad (1)$$

b) Model-based Collaborative Filtering: For experiments with collabotive filtering method, we select Species_FDIR_code, Location_code and Rating as input for Alternating Least Squares (ALS) algorithm in order to learn latent factors that can be used to predict missing entries of user-item association matrix.

We implement built-in ALS model supported in spark.ml with following parameters: maxIter = 20, rank = 200, regParam = 0.04 for our experiments. Once the model is set up, we fit it on train set and predict ratings in test set to achieve the result.

c) Hybrid Recommendation: To create hybrid recommendation system, a weighting technique is applied to combine different approaches. In this report, the incorporation of Content-based Filtering (CB) and Collaborative Filtering (CF) is based on the Equation 2. By trying different weights σ combining two methods and comparing the results with the average combination, the optimized performance is achieved with $\sigma = 0.45$.

$$Hybrid = \sigma \cdot CF + (1 - \sigma) \cdot CB \quad (2)$$

B. Species Gross Weight Forescasting

1) Features Selection and Preprocessing:

Since haddock and mackerel are two species of our interest to forecast gross weight, we create two subset of catch notes data by filtering each species. The features we choose from the dataset for experiments is the last catch day of vessels before landing and indispensably the gross weight of the fish. Besides historical data, we also use the sea surface temperature (SST) from NOAA SST dataset and seven datetime features generated from timestamp including: DAY, DAYOFYEAR, WEEKDAY, WEEKOFYEAR, MONTH, YEAR and IS_WEEKEND in our experiments. The sea temperature data is stored under netCDF4 format and is easily accessible using netCDF4 Python module.

Finally, since there many catches take place at the same time, we groupby the timestamp and calculate the average gross weight of the species. The final output of our models is the daily average gross weight of either haddock or mackerel. For training and evaluation, the subsets are split into train set and test set with ratio of 9:1 respectively. The period of test set roughly start from April 2020 for haddock, July 2020 for mackerel and both end in 05, May 2022.

2) Experiments Setup:

We implement two built-in time series forecasting models including **TCNForecaster** and **Seq2SeqForecaster** supported in Chronos library of BigDL framework. These models are set up with the same default hyperparameters for two fishes at first, then they will be separately adjusted and fine-tuned to get the better results. Training configurations such as epoch and batch size are constantly kept throughout the experiments. The experiments are carried out using different combinations of features respectively on forecasting gross weight of each fishes to identify the best combinations and study their effects. Those combination of features are scaled to have zero mean and unit variance before feeding into models. The two models are configured to give the prediction up to four days based on data of the last 90 days for each type of fish.

C. Experimental Results

1) Fishing Locations Recommendation:

The performance of fishing location recommendation system are evaluated in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and shown in Table IV. Content-based filtering method gives better result than ALS model with RMSE of 0.5228. Weighted hybrid recommendation, which is the combination of content-based filtering and collaborative filtering, achieves the best performance among methods with RMSE of 0.4933, better than average combination 0.0007 rating value. Some examples of fishing location suggested by ALS models are shown in Table V. In the table, four fishing locations represented by ids with the hishest rating are recommended for catching each type of fish species.

TABLE IV
RESULTS OF FISHING LOCATION RECOMMENDATION

Method	RMSE	MAE
Content-based Filtering (CB)	0.5228	0.3493
ALS	0.5434	0.3688
Average (CB + ALS)	0.4940	0.3317
Weighted (CB + ALS)	**0.4933**	**0.3311**

TABLE V
FISHING LOCATIONS RECOMMENDED BY ALS MODEL.

Species Name	**Recommended Fishing Locations**
Mackerel	[{**83**, 4.21}, {**81**, 4.13}, {**79**, 4.12}, {**59**, 4.04}]
Haddock	[{**58**, 2.70}, {**69**, 2.37}, {**45**, 2.31}, {**78**, 2.17}]
Capelin	[{**59**, 5.53}, {**58**, 5.23}, {**05**, 4.99}, {**45**, 4.97}]
Skagerak Herring	[{**29**, 4.83}, {**06**, 4.11}, {**14**, 4.06}, {**09**, 4.06}]

2) Species Gross Weight Forecasting:

The RMSE results for species gross weight forecasting using TCN and Seq2Seq model are shown in Table VI. For haddock, TCN model achieves best result forecasting one day ahead with RMSE 1,807.46. While for mackerel, Seq2Seq model perform better with RMSE of 76,743.82. However, when predicting the next two to four days, TCN seems to have better results with minimum RMSE of 73,668.13 forecasting 4 days ahead.

The usage of additional features gives significant improvement to the performance of both models. Among those features, it can be seen that datetime features are important factors when 7/16 best performances are achieved using them in experiments. The results also show that the combination of datetime features and sea surface temperature (SST) together helps improve the performance of Seq2Seq in all four days forcasting makerel's gross weight.

Comparative illustrations of the predicted gross weight one day ahead for each species given by two models on test set are described in Figure 4 and Figure 5. For haddock, it seems that prediction graph (red) of both models do not fit with the ground truth (blue) very well since the gross weight of haddock have high variance between days. However, we can see that TCN perform better when it recognizes the trend in increasing gross weight around March, April and May each year. For mackerel, the prediction graph made by Seq2Seq model has a good fit with the ground truth, while prediction graph made by TCN model just simply covers the shape of the ground truth, misses some small trends and even gives negative gross weight for some periods.

VI. CONCLUSION

Recommendations for fishing locations and gross weight forecasts are essential for sustainable fishing development. Both of these two tasks will shorten the searching period, which is translated to CO2 emission and operation cost reduction and in overall will increase the profitability of fishery industry.

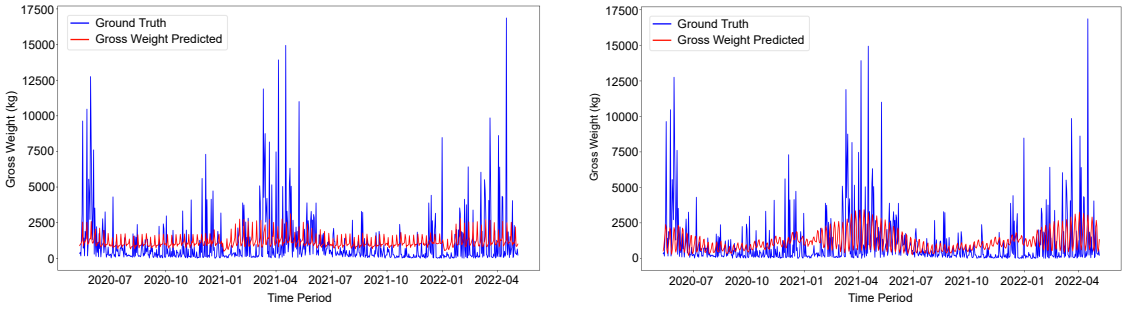
This study proposes approaches using big data frameworks and applications to solve realistic problems toward sustanable commercial fishing. We have built a hybrid recommendation system model for suggesting fishing locations to optimize the catching of specific fishes and experimented various deep learning based multi-variate time series models to forecast the gross weight of 2 important species of Norwegain fishery: haddock and makerel. To ensure the practicality of the application, all models are trained and deployed using big data technologies. The results show that the both proposed hybrid recommendation system model and time series forecasting models achieve high performance regarding suggesting fishing location and predicting fishing output.

ACKNOWLEDGMENT

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

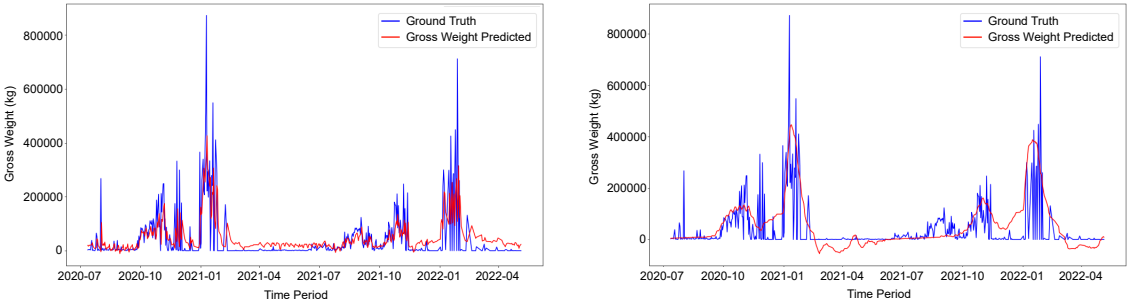
TABLE VI
THE RMSE RESULTS COMPARE FORECASTING MODELS PREDICTING GROSS WEIGHT AVERAGE OF HADDOCK AND MACKEREL

Species	Extra Features	TCN				Seq2Seq			
		1 day	2 days	3 days	4 days	1 day	2 days	3 days	4 days
Haddock	None	1,837.71	1,877.56	**1,834.28**	1,856.19	1,890.48	1,906.95	**1,856.43**	1,851.15
	Datetime	**1,807.46**	1,922.03	1,867.84	**1,806.20**	**1,853.50**	**1,807.43**	1,866.42	1,891.24
	SST	1,816.58	**1,855.24**	1,941.72	1,942.41	1,867.60	1,856.33	1,867.40	**1,850.02**
	Datetime + SST	1,819.89	1,897.96	1,921.80	1,830.79	1,861.59	1,843.33	1,861.74	1,877.19
Mackerel	None	87,246.46	75,759.24	75,027.06	79,129.95	77,254.79	77,550.57	77,742.23	79,125.18
	Datetime	**82,953.99**	74,654.14	**74,032.45**	**73,668.13**	79,050.70	79,034.90	78,371.35	78,995.88
	SST	105,219.27	**74,598.79**	79,531.50	91,694.21	77,816.29	78,215.28	79,275.61	80,525.32
	Datetime + SST	84,727.09	81,293.14	76,093.35	80,100.58	**76,743.82**	**76,666.55**	**76,605.64**	**77,715.60**



(a) Seq2Seq model (b) TCN model

Fig. 4. Haddock's gross weight one day forecasting using Seq2Seq and TCN



(a) Seq2Seq model (b) TCN model

Fig. 5. Mackerel's gross weight one day forecasting using Seq2Seq and TCN

REFERENCES

[1] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, dec 2015. [Online]. Available: https://doi.org/10.1145/2827872

[2] J. Bennett and S. Lanning, "The netflix prize," in *Proceedings of the KDD Cup Workshop 2007*. New York: ACM, Aug. 2007, pp. 3–6.

[3] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 22–32.

[4] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, *Content-Based Movie Recommendation System Using Genre Correlation: Proceedings of the Second International Conference on SCI 2018, Volume 2*, 01 2019, pp. 391–397.

[5] M. Goyal, "Collaborative filtering movie recommendation system," *International Journal for Modern Trends in Science and Technology*, vol. 6, pp. 471–473, 01 2021.

[6] A. Singh and P. Soundarabai, "Collaborative filtering in movie recommendation system based on rating and genre," *IJARCCE*, vol. 6, pp. 465–467, 03 2017.

[7] S. Suriati, M. Dwiastuti, and T. Tulus, "Weighted hybrid technique for recommender system," *Journal of Physics: Conference Series*, vol. 930, p. 012050, 12 2017.

[8] M. Ospici, K. Sys, and S. Guegan-Marat, "Prediction of fish location by combining fisheries data and sea bottom temperature forecasting," in *Image Analysis and Processing – ICIAP 2022*. Cham: Springer International Publishing, 2022, pp. 437–448.

[9] M. Iiyama, K. Zhao, A. Hashimoto, H. Kasahara, and M. Minoh, "Fishing spot prediction by sea temperature pattern learning," in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018, pp. 1–4.