Contribution of Collaborative Filtering Approach on Environmental Big Data Analytics Context

ABAROU Salma, EL ABDERRAHMANI Abdellatif and SATORI Khalid

LISAC, Department of Computer Sciences Faculty of Sciences DHAR EL MEHRAZ University Sidi Mohamed Ibn Abdellah

Abstract. In recent years and following the spread of internet using, the way of life in our society as well as the management of the main production sectors have changed towards digitalization (E-Commerce, E-Surgery, road traffic management, etc.) to minimize the . This change challenged researchers and organizations to come up with solutions to manage, maintain, classify and make the best decision from the gigantic flows of data created daily while reducing environmental Damage (Consumption of Energical resources….) ,From which was born the Collaborative filtering approach which facilitate decision making based on the experience feedback of Internet users. This approach is used generally the big Data Analytics Algorithm to predict and classify data

1 Introduction

Due to the rise of internet usage, users have daily an endless supply of information about various topics (news, choice of products on E-Commerce Websites…etc.).
These massive amounts of data flows put researchers and organizations in a difficult situation as they try to innovate solutions that help internet users make best decisions while minimizing unnecessary data from incoming flows to protect the environment (.
Hence the birth of the concept "Collaborative Filtering" [1] which is a set of algorithms aimed at building a System of recommendation to an individual based on the assessments of a group of people having a similar profile. This technique is spread over several approaches such as the User Model Approach which consists to build models of users based on their information and the Item-Centered Approach consists to use a measure of similarity between items to determine the K items most similar to item A.
This paper presents a comparative study of various collaborative filtering algorithms that use big data Analytics Algorithm. It is structures as follows Section 2 describe Collaborative Filtering Concept and the main families of this method, While Section 3 present the Decision Tree Algorithm [6] by taking as an example the CHAID Method [5]. Section 4 presents the K-NN Algorithm [2-4] Section 5 represents an interpretation and finally Section 6 is conclusion of our work.

2 Collaborative Filtering Approach

2.1. Presentation

In general, we can say that Collaborative Filtering [8] is a set of techniques of making automatic predictions (Filtering) about User interest by collecting preferences of many users (collaborating).
This technique is used by Recommender System -to filter data sets by techniques involving collaboration among multiple agents, viewpoints, data sources- on different kinds of data including (Monitoring Data, Financial Data, and Electronic Commerce Data).
This process is based on the following steps:

- Firstly, System collect all User Information to obtain their preferences, this collect can be
  - Explicit: User assign rating to their products or they indicate their appreciation for example by Like.
  - Implicit: the collect is based on behaviors (purchase, click and duration on a page).
- Secondly, System compare New User information with that of similar users.
- Finally, system recommended the products to user based on the similarities founded

☒ Today, several Web-Sites and organization (YouTube, Jumia, Netflix…) are using this approach to helps their user find the most suitable products.

2.2. Main families of Collaborative filtering approach

Nowadays, we can distinguish between three main families of Collaborative filtering approach:

Table 1. Collaborative Filtering Approach

Filtering System	Description

* Corresponding author: salma.abarou2014@gmail.com

Active Filtering System	Use the explicit tastes of users (this is thanks to a rating system) on a range of products
Passive Filtering System	Analyze the history of a user's behavior to deduce their preferences.
Filtering Based on the content System	Based on the characterization of the content of the information to be filtered (i.e. the information is only retrieved if it coincides with the topics of interest to the user).

- Today, many Big Data Analytics Algorithms are used on Collaborative Filtering approach. In our study, we used two types of algorithms given their power and reduced consumption of resources

3 Decision Tree Algorithm

3.1. Presentation

The decision Tree [6, 10] is a decision support tool which represent a set of decision in graphical form of Tree. As the final decision are located at the end of branch (leaves) and they are reached by the choices made at each steps.

- This method makes it possible to distribute the elements of a heterogeneous groups according to a set of discriminating variables and according to a fixed objective.
- It is generally used in various fields such as the banking Sector (Processing Credit Request from a set of customers), Medicine (Preventing adverse effects of a medical care) and Psychology Sector (study of models human behavior).

Table 2. Decision Tree Algorithm

Decision Tree Algorithm	**Description**
Classification And Regression Trees Algorithm (CART Algorithm)	Algorithm with the aim of classifying a set of records according to the segmentation criterion "The Gini index"
Quick Unbiased Efficient Classification Tree (QUEST method)	Algorithm allowing to explain a qualitative variable having a large number of modalities
Chi-squared Automatic Interaction Detector 'CHAID Algorithm'	Technique used for the detection of interaction between variables using the criterion' Chi-square formula'

3.2. Advantages and Disadvantages

Nowadays, the popularity of decision tree method is based on the fact that this method is easy to understand and allows to increment already existing trees by new options and to choose the most appropriate option and finally we can easily combine it with other decision-making tools.
Despite these advantages, this method has certain limitations such that in some cases one can have very complex trees (to remedy this concern of over learning, new pruning procedures are used), and the absence of the expression of certain concepts in decision trees (XOR, parity….).

3.3. Principle of the method

In this part, we have chosen to put Decision Tree Method into practice by studying the Chaid Decision Tree Method [5] which generates no-binary trees (i.e. a node can have several branch)

- This method can be applied to all types of inputs and accepts observation weights.
- To establish a chaid type decision tree, sur must respect the following order:
  - Define the root of the decision tree
  - Choose the segmentation variables using the KHI-2 formula
  - Identify the nodes and leaves of the tree.

REPEAT
Taking into account the vertex to segment
Preparation of quantitative variables (discretization, choice of a cut-off)
Choose the best segmentation variable (by using index)

IF the selected variable is qualitative **THEN**
Test of fusion of modalities with a similar profile
Merge if the tests are significant
END-IF
UNTIL THE Stop Condition

- The KHI-2 formula is a technique founded by statistician Karl Pearson in 1900 which makes it possible to test the adequacy of a series of data to a family of probability laws or to test the independence between two random variables.

$$x^2 = n(\sum_{i,j} \frac{(n_{i,j})^2}{n_i . n_j} - 1)$$

Such as

- $n_{ij:}$ is the number contained in the box spotted by the i line and j column.
- n: Total workforce.

- To understand the application of this formula, we decided to use the following example to predict the possibility of playing a tennis match depending on the weather.

Table 3 : Data Set Of CHAID Algorithm

N°	sunshine	temperature (°F)	Humidity (%)	Play
1	Sun	75	70	YES

2	Sun	80	90	NO
3	Sun	85	85	NO
4	Sun	72	95	NO
5	Sun	69	70	YES
6	covered	72	90	YES
7	covered	83	78	YES
8	covered	64	65	YES
9	covered	81	75	YES
10	Rain	71	80	NO
11	Rain	65	70	NO
12	Rain	75	80	YES
13	Rain	68	80	YES
14	Rain	70	96	YES

- The first step is to define the frequency , the possibility of the game based on the Sun Factor

Table 4 The frequency and the possibility of the game

Sunshine/Play	**YES**	**NO**	**Total**
Rain	3	2	5
Covered	4	0	4
Sun	2	3	5
Total	9	5	14

- The second Step is based on the calculation of the T0 value for all cases

$$T_0(\text{Rain, YES}) = \frac{5*9}{14}\frac{45}{14}$$
$$T_0(\text{Rain, NO}) = \frac{5*5}{14}\frac{25}{14}$$
$$T_0(\text{Covered, YES}) = \frac{4*9}{14}\frac{18}{7}$$
$$T_0(\text{Covered, NO}) = \frac{4*5}{14}\frac{10}{7}$$
$$T_0(\text{Sun, YES}) = \frac{5*9}{14}\frac{45}{14}$$
$$T_0(\text{Sun, NO}) = \frac{5*5}{14}\frac{25}{14}$$

Table 5. Calculation of the T0 value

Sunshine/Play	**YES**	**NO**
Rain	3,21	1,78
Covered	2,57	1,42
Sun	3,21	1,78

- The third step aims to develop the table of deviations from independence $R^2 = (T\text{-}T_0)^2$

Table 6 . Calculation of R2 = (T -T0°)2

Sunshine/Play	**YES**	**NO**	**Total**
Rain	0,05	0,05	0,09
Covered	2,04	2,04	4,08
Sun	1,47	0,83	2,30
Total	3,56	2,91	6,47

- In on our study , we use Sipina Software to implement the decision Tree Chaid based on a data set to predict if there is a relation between the frequency of fire , and the high temperature , the flow of water (we use 2942 rows) , we obtain an simple hierarchical result on 1200 ms.

- To conclude this section, we can say that the decision tree method is generally used in areas of decision support, or data mining. It makes it possible to distribute a population of individuals into homogeneous groups according to a set of discriminating variables and according to a fixed objective.

4 Nearest Neighbors Algorithm

4.1. Presentation

The K-NN Method [2-4] is a supervised learning algorithm aims to classify objects according to their similarities with the object in the learning base.

- In a context of classification of a new classification of a new observation x, the simple founding idea is to have the nearest neighbors vote on this observation. The class of x is determined according to the majority class among the k closest neighbors to the observation x.
- This method is generally used in various fields such as image processing (Pattern recognition process), Medical diagnosis (Detection of biomarkers during medical screenings), astronomy (Analysis of images received by satellites) and finally e-commerce (targeted marketing based on customer opinions).

4.2. Advantages and Disadvantages

Nowadays, the popularity of the K-NN method is based on the fact that this method is easy to understand and adapted to the fields where each class is represented by several prototypes and whose borders are irregular (such as the recognition of handwritten figures or satellite images).
Despite these advantages, this method has certain limitations such as slow prediction since each time it reviews all of the examples, the method takes up a lot of memory.

4.3. Principle of the K-NN method

The principle of K-NN Algorithm is based on the following steps:

- Define the input data (the learning base D, choice of the distance function d, the parameter k and finally the data of unknown class x).
- The data of unknown class x is compared with all the data stored in the learning base d. We

choose the majority class among its k closest neighbors in the sense of a chosen distance.

```
BEGIN
FOR any element a belonging to D DO
        Calculate the distance dist (x, a)
END FOR
FOR any element a belonging to kppv(x) DO
        Count the number of occurrences of each
    class
END FOR
        Assign x to the nearest class
```

☒ We can distinguish between different distance calculation methods, this table presents the most used:

Table 7 Distance Calculation Method

Distance Calculation Method	**Description**
Manhattan Distance	The route taken by a taxi when it moves from one node of the network to another using the horizontal and vertical movements of the network. $d(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$ ☒ It is used to calculate the sum of the absolute values of the differences between the coordinates of two points.
Euclidean distance	The shortest distance between two points, it is calculated by the following formula $d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$ ☒ this method is used for quantitative data having the same type
Minkowski Distance	The measurement in a normalized vector space which can be considered as a generalization of both the Euclidean distance and the distance from Manhattan. $d(x,y)=\sqrt[q]{\sum_{i=1}^{n}\lvert x_i-y_i\rvert^q}$

☒ To understand application of calculation methods, we decided to base on the following example which makes it possible to predict if 'Sebastian' is loyal to a bank

Client	**Age**	**Revenue**	**N° of Credit Cards**	**Loyal**
John	35	35K	3	NO
Rachel	22	50K	2	YES
Xavier	63	200K	1	NO
Anthony	59	170K	1	NO
Nellie	25	40K	4	Yes
Sébastien	37	50K	2	?

- The first step is to choose the parameter k, for our case we choose the value 3.
- The second step consists in calculating the distance, for our case we chose the Euclidean distance:

The distance between Sebastian and 3-NN
Sqrt [(35-37)²+ (35-50)²+ (3-2)²]=15,16
Sqrt [(22-37)²+ (50-50)²+ (2-2)²]=15
Sqrt [(63-37)²+ (200-50)²+ (1-2)²]=152,23
Sqrt [(59-37)²+ (170-50)²+ (1-2)²]=122
Sqrt [(25-37)²+ (40-50)²+ (4-2)²]=15,47

☒ We can say that Sébastien is a loyal client.
☒ In our study we use TANAGRA Software to process 3442 rows of data set to predict if there is a relation between the frequency of fire , and the high temperature , the flow of water (we use 3442 rows) . The result are Computation Time is 797 ms an allocated memory is 252KB.

5 Interpretation:

Following our study, we can say that two main algorithms of the Collaborative Filtering vision are decision trees and the K-NN algorithm.
Indeed, the Decision Tree method makes it possible to distribute the elements of a heterogeneous group into homogeneous groups according to a set of discriminating variables and according to a fixed objective.
In return, the K-NN method is a collaborative filtering technique which aims to find the k closest neighbors for each of the objects according to a distance calculation method. After finding the neighborhood closest to the object, a majority vote is taken to make the recommendations.

6 Conclusion

This article represents a synthesis of the comparative study between the two main algorithms of the Collaborative Filtering vision, namely decision trees and the K-NN algorithm.
Given its characteristics (the K-NN algorithm consumed 797ms to trait the data, the Chaid algorithm consumed about 1200 ms to treat the same data) and their results, we can say that the most suitable algorithm for developing active recommendation systems is the K-NN algorithm.
Our goal is to improve the K-NN algorithm to reduce processing time and resource consumption thus ensuring data security by adding the implementation of security algorithms.

References

1. F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma. Collaborative knowledge base embedding for recommender systems. In KDD, pages 353--362, (2016).
2. Eve Mathieu-Dupas, Algorithme des k plus proches voisins pondérés ET application en diagnostic (2010)
3. Hechenbichler, K. ET Schliep K. Weighted k-nearest-neighbor techniques and ordinal Classification. Sonderforschungsbereich 386, Paper 399(2004).
4. Salvador–Meneses, J.; Ruiz–Chavez, Z.; Garcia–Rodriguez, J. Compressed kNN: K-Nearest Neighbors with Data Compression. Entropy 2019, 21, 234. (2019)
5. M. Ondokuz, H. ÖNDER, Use of CART and CHAID Algorithms in Karayaka Sheep Breeding (2019)
6. Maji S., Arora S. Decision Tree Algorithms for Prediction of Heart Disease. In: Fong S., Akashe S., Mahalle P. (eds) Information and Communication Technology for Competitive Strategies. Lecture Notes in Networks and Systems, vol 40. Springer, Singapore (2019)
7. Jiang, L., Cheng, Y., Yang, L. et al. A trust-based collaborative filtering algorithm for E-commerce recommendation system. J Ambient Intell Human Comput 10, 3023–3034 (2019).
8. A.Murat Turk, A.Bilge, Robustness analysis of multi-criteria collaborative filtering algorithms against shilling attacks, Expert Systems with Applications, Volume 115, Pages 386-402(2019)
9. Hong, B., Yu, M. A collaborative filtering algorithm based on correlation coefficient. Neural Comput & Applica 31, 8317–8326 (2019)
10. Li, Y., Jiang, Z.L., Yao, L. et al. Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. Cluster Comput 22, 1581–1593 (2019)