PSO-FSDP Clustering for Internet Propagation

Yu Fangting
Universiti Teknologi Malaysia
Johor Bahru, Malaysia, Malaysia
Werewrwerw34@protonmail.com

ABSTRACT: *This article studies the PSO-FSDP clustering problem based on internet propagation. Much data needs to be analyzed and classified in the internet environment to understand better and grasp important information, such as user behavior and market trends. Traditional clustering algorithms often face problems such as low computational efficiency and inaccurate results when dealing with large-scale and high-dimensional data. Through experimental verification, the PSO-FSDP algorithm has higher accuracy and stability when processing large-scale and high-dimensional internet data and has faster computational speed than traditional clustering algorithms. Therefore, the PSO-FSDP algorithm provides an effective method for solving clustering problems in internet propagation and has important theoretical significance and practical value.*

Keywords: Topic Discovery, PSO-FSDP Clustering Algorithm, News Analysis

Received: 4 March 2023, Revised 22 May 2023, Accepted 22 June 2023

DOI: 10.6025/ijwa/2023/15/3/73-79

1. Introduction

With the rapid development of Internet technology, we are in an era of information development with exponential growth. Now, the Internet has become one of the main ways for the public to get information, unlike the radio and newspaper era [1]. Now, we are facing an era of information overload, which has become the primary problem for their use of information in so much information; the information of Internet news is complex, as one of the most important and the number is extensive information [2]. In the statistical result of the development of the Internet Network Center in China, the number of netizens in our country has reached 60% of the total population [3].

When a hot news event occurs, there will be a lot of news websites for different types of it in a release. The number of news websites where people receive news information often exceeds the understanding of their absorption and for news and information, the traditional search engine can be improved on this phenomenon to a certain extent [4]. However, there are some limitations in the processing effect brought by search engines. In the face of massive news information, we need better news processing technology to extract information quickly. So, we need to detect the news hotspots through the basic algorithm technology. After processing, we can quickly understand the current hot news [5].

2. State of the Art

Topic detection technology was first used in the topic detection and tracking field; both use similar techniques. TDT is a new technology developed by the United States Department of Defense and the National Institute of Standards and Technology; its purpose is to extract all kinds of information classification information processing to achieve the processing of the information overload problem so it can effectively carry out effective processing of news information [6]. So this time, we use the topic detection technology and then combine the clustering algorithm to extract news information. Research in this way is bound to be rewarding. Moreover, as time passed, our country began introducing this technology in 1999. Li Baoli and Oracle, two of Peking University, studied detection and tracking technology [7]. The related prediction and analysis of the current research status and future development trend in the TDT field [8].

At present, the topic detection system is mainly used in the micro-blog and BBS website; the primary function is for micro-blog and BBS forum user emotion analysis and then push messages, but for Web hot news discovery is less, how to find a balance of efficiency and accuracy it is very important [9]. We need to apply the existing clustering algorithm to the online clustering and algorithm improvement of news reports because this can enhance the efficient application of the algorithm. This can be a good search for hot news [10].

3. Methodology

3.1. Clustering algorithm Model for News Topic Discovery

Before the algorithm model, we need to have the information extraction and classification because this way can allow us to fast computation and processing process in the construction of the calculation model of the algorithm so that we will use the text by a vector space model for recording and classification. The Boolean model is a typical set of Boolean models; in the model building, the first step is to extract the key information of text keywords by defining a series of two-element feature vectors. And then use the features extracted before the representation of this article:

$$D_1 = (d_1, d_2 K, d_n) \tag{1}$$

In the upper form, n is the number of the characteristic variables, and d is the probability that appears in the text and is assigned to it by the pair. The occurrence of the assignment is True, and the error is set to False. Then a non-two-element method is used to represent the weight of the feature by establishing a space vector:

$$V(d) = (t_1, W_1(d); t_2, W_2, (d); K\, t_m, W_m, (d)) \tag{2}$$

By type, characteristic values of the mean vector are characterized by the use of text representation, the type t as feature sets of documents in d, w as the weight of words, through a representation of feature weight, can get to statistical classification for each piece of information. Through such a topic model, it is possible to generate random documents in the way of the polynomial distribution of the topic. The topic determines every word in each document, and the feature words produced by each topic are also different. The next is to calculate the similarity of the research object of the data, which is an important part of the PSO-FSDP clustering algorithm. The principle results of the clustering algorithm are shown as shown in Figure 1:

The choice of similarity calculation is through the following two different vectors, of which n is the dimension of two vectors.

$$d_1 = (a_1, a_2 K, a_n) \tag{3}$$

$$d_2 = (b_1, b_2 K, b_n) \tag{4}$$

The formula for calculating the similarity between them is as follows:

- **The inner product similarity:**

$$sim\,(d_1, d_2) = d_1 \bullet d_2 = \sum_{i=1}^{n} (a_i \times b_i) \tag{5}$$

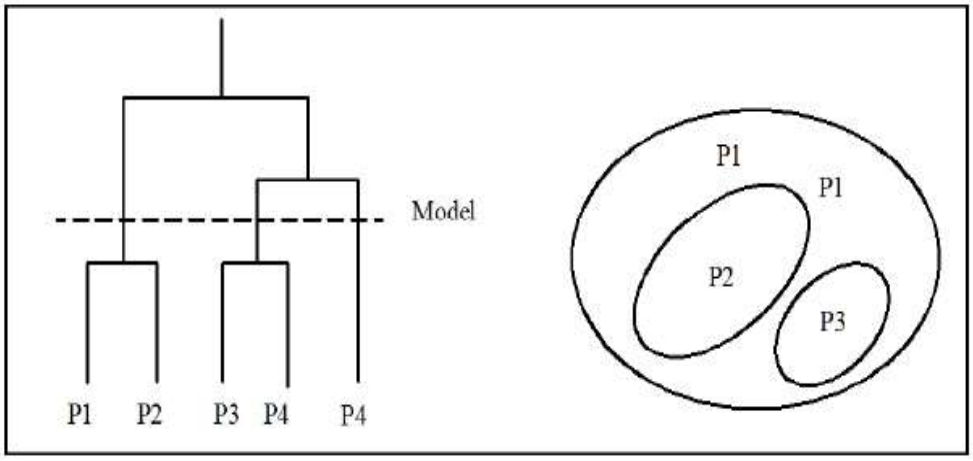

Figure 1. Management model of clustering algorithm schematic image as an effect diagram

- **cosine similarity:**

$$sim\ (d_1, d_2) = \frac{d_1 \bullet d_2}{d_1 \times d_2} = \frac{\sum_{i=1}^{n} (a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} a_i^2 \times \sum_{i=1}^{n} b_i^2}} \tag{6}$$

- **Jaccard similarity:**

$$sim\ (d_1, d_2) = \frac{d_1 \bullet d_2}{|d_1| \times |d_2|} = \frac{\sum_{i=1}^{n} (a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} a_i^2 \times \sum_{i=1}^{n} b_i^2}} \tag{7}$$

By calculating the upper form, we collect the feature values of all news information and then use the similarity algorithm to calculate the similarity of the information obtained and classify the news information. In this way, each information acquisition can reduce the environment of information acquisition, and in this way, we can divide a set of information data. As the saying goes: “Like attracts like. Birds of a feather flock together.” In natural science and social science, there are many classification problems. The data of the algorithm are recorded as shown in Table 1 as shown below:

Generally speaking, the so-called class refers to the collection of similar elements. Cluster analysis is based on taxonomy. In ancient taxonomy, people mainly relied on experience and professional knowledge to classify them and rarely used mathematical tools for quantitative classification. With the development of human science and technology, the demand for classification is increasing, so it is difficult to classify accurately by experience and professional knowledge. So, people gradually introduced mathematical tools to taxonomy and formed numerical taxonomy. Then, the technology of multivariate analysis was introduced to numerical taxonomy, and cluster analysis was formed. The content of cluster analysis is very rich. There are systematic clustering methods, ordered sample clustering methods, dynamic clustering methods, fuzzy clustering methods, graph theory clustering methods, clustering prediction methods and so on. In business, clustering can help market analysts distinguish different consumer groups from the consumer database and summarize each consumer’s consumption patterns or habits. As a module in data mining, it can be used as a separate tool to find out some deep information distributed in the database and generalize the characteristics of each class or to focus on a specific class for further analysis. Clustering analysis can also be used as a preprocessing step for other analysis algorithms in data mining algorithms.

Initial data value	Threshold of logical unit	Expected response value	The characteristic vectors are computed
100	65	65	15%
200	120	120	14%
300	180	179	16%
400	225	205	14%
500	303	300	22%

Table 1. The Coefficients of the Data of this Algorithm are Recorded as Follows

3.2. News Topic Discovery based on PSO-FSDP Clustering Algorithm

Through the algorithm formula above, we can know that the parameters involved in the above formula are also very few because the algorithm is relatively simple; it also allows us to be more accurate and more convenient in operation. To make our algorithm more stable, we need to perform the standard function for the maximum speed V, so we express the formula in the position of the best solution of the global algorithm.

$$f_{\sigma}(\chi) = 0.5 + \frac{\left(Sin\sqrt{\chi^2 + y^2} \right) - 0.5}{\left(1.0+0.001(\chi^2 + y^2)\right)^2} \tag{8}$$

In the upper form, we can see that the inertia weight factor w and the maximum speed V of our algorithm calculate all the simulation processes. At present, we set the number of populations as subgroup 20. We can revise and decorate it and get the calculation and setting of the inertia weight factor at different computing speeds. We inputted the fuzzy inference mechanism into three sets in the domain. This time, we use a definition of the Triangle membership function.

$$f_{Triangle}(\chi) = \left\{ \begin{array}{lll} 0 & if & \chi < \chi_1 \\ 2 \times \dfrac{\chi_2 - \chi}{\chi_2 - \chi_1} & if & \chi_1 \le \chi \le \dfrac{\chi_2 + \chi_1}{2} \\ 0 & if & \chi > \chi_2 \end{array} \right\} \tag{9}$$

The membership function in the upper part defines three different parts of x, and gets the data value and determines the increment of the current w. We calculate the best optimum value through a lot of data for these benchmark functions. We take into account the efficiency and search accuracy of the search. As the most effective search method, we will adaptively adjust a global search method. Based on this, we have evolved an evolutionary function of our current algorithm through the idea of genetic algorithms. We combine the selection mechanism with the PSO-FSDP algorithm, and we can compile the iterative algorithm for the survival of the fittest. The selection mechanism of the hybrid PSO-FSDP algorithm and the other evolutionary optimization algorithms are most suitable for each individual's current location. The probability of the hybridization given by the particle in the particle swarm is determined by the user of our algorithm. We select the specified number of particles according to the probability of crossing into a natural hybridization environment.

Through the large number of the population between the premise of the same particle, we can add the algorithm; the calculation formula is as follows:

$$child(X) = parent(X) + (1 - p)\, parent_2(X) \tag{10}$$

X is the location vector of D, where the child and parent are selected as the parent-offspring in the natural data exchange location. The components of the crossover probability are all between 0 and 1. Big data brings us three subversive ideas: all data, not random sampling; general direction rather than precise guidance; correlation rather than causality. Not a random sample, but a whole data. In the era of big data, we can analyze more data, sometimes even process all data related to a particular phenomenon, and no longer rely on random sampling. (We usually regard random sampling as a necessary limitation, but high-performance digital technology makes us realize this is a human restriction.). Not accuracy, but confounding: There is so much research data that we are no longer keen on accuracy. There are few data to be analyzed before. Therefore, we must quantify our records as accurately as possible. With the expansion of scale, the obsession with accuracy will be weakened. With big data, we no longer need to get to the bottom of a phenomenon; as long as the master of the general development direction can be ignored, the appropriate accuracy on the micro level will let us have a better insight into the macro level; not causation, but correlation: We are no longer keen on causality, and finding causation is a human habit for a long time. In the era of big data, we don't need to stare at the causality between things, but we should find out their relationship. The relationship may not tell us exactly why something will happen, but it will remind us that this is happening. A schematic diagram of the running process of the algorithm is shown in Figure 2 below:

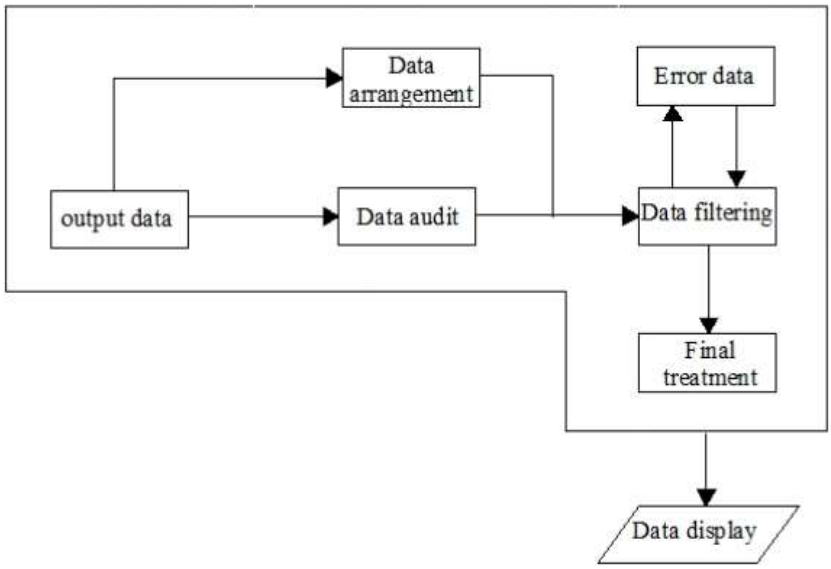

Figure 2. A schematic diagram of the detailed calculation process of this research algorithm

4. Result Analysis and Discussion

We need to detect the algorithm after the news topic discovery research algorithm based on the PSO-FSDP clustering algorithm completes the modeling. To study the time to verify the feasibility of the algorithm, we will use the PSO-FSDP clustering algorithm research model according to the network data environment for optimal selection, as the era of big data is mixed because there are a lot of data which is in people's daily life and production activities, including a lot of useless information. We need to find news information that we need in this complicated information to achieve our value in the current era. The most essential source of the PSO-FSDP clustering algorithm is to study the classification and tracking of news information. According to the classic test of the classic test problem-solving problem of the PSO-FSDP clustering algorithm, we use the method of the Rosenbrock function test. Rosenbrock is a function of a single peak value; the variables between us are strong. Its overall situation is distributed in a tiny area, and it is distributed between 0 and 1. On this basis, the Griewark function is used as a function of multiple peaks that influence each other, and its locality is also at some minimal points. The calculation records are shown in Table 2 as follows:

		CPSO				SPSO	SGA
Function	Dim	W	C1	C2	Fitness	Fitness	Fitness
Rosenbrock	10	0.659	2.796	1.335	0.159	3.214	10.452
Griewank	10	0.612	2.755	1.256	0.035	0.085	0.108

Table 2. Algorithm Overall Performance Test Results Summary Record Control Table

From the above table, we can see that the crossover rate of our SGA is p = 0.2, and the optimal solution of group n=30 is achieved after many times running, so the SPSO optimal value is better than the previous algorithm. So, the algorithm of our study has great superiority. The result is that the SPSO optimization result is better than the SGA. So, we use the PSO-FSDP clustering algorithm to meet the news classification and obtain information. The method of SPSO and CPSO is used to calculate the parameters. Compared with the SGA method, the selection of the control parameters and the test of the standard function are the same. So in order to avoid the contingency of the test, the method of cross validation is used for the estimation and the test of the result. Then, we tested 60 of the samples we encountered. We have 30 run tests on each group of data, as shown in Table 3 as follows:

Data value	Test value of the algorithm	Parameter values of Technology	Result accuracy
10	30%	70%	95%
20	45%	90%	90%
30	38%	98%	92%

Table 3. The Running Record Table of Multiple Data Calculation of Group Data

From the previous table, we can see that the algorithm model has great stability and convenience when it is inputting large data.

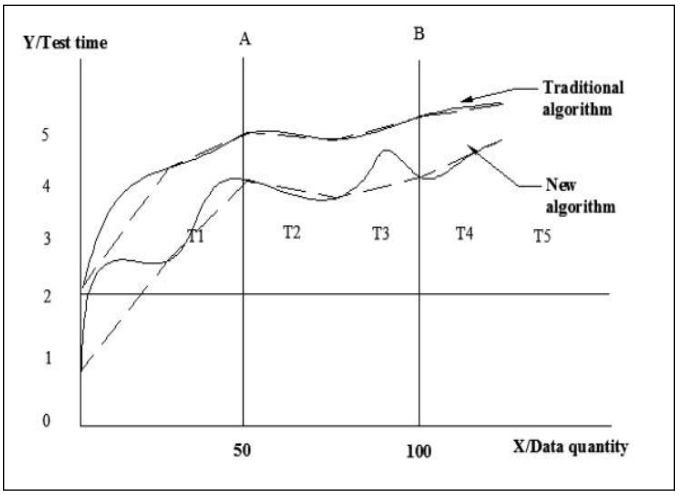

Figure 3. The performance comparison diagram of the improved clustering algorithm and traditional clustering algorithm

Our news topic discovery based on the PSO-FSDP clustering algorithm has greatly improved the computing power of the previous algorithm model. This is a faster way to perform our data operation and a more accurate and practical design scheme in the practical application. A new approach to researching JAVA's multi-thread computer language is adopted. We compare the previous clustering algorithms and contrast, as shown in Figure 3:

From the figure 3, we can see that the algorithm of this study is entirely in line with our current network information age model. We can quickly find the information we need through the algorithm model and then combine our information based on it. So, this research algorithm can meet our needs and is a vast improvement in the algorithm.

5. Conclusion

With the rapid development of Internet technology, the Internet as a new media and the three traditional media are called the four largest media. It has become one of the main ways for the public to get information, publish messages and transmit messages in the current society. There are also a lot of news media in today's Internet platform, the Internet news has become a "sharp pioneer" of news reports. A lot of information is spread on the Internet every day. Research on the hot topics of the news has become an essential job in every news industry. In this paper, Research on news topic discovery is based on the PSO-FSDP clustering algorithm. The domestic topic discovery system research is mainly used in the major micro-blog and BBS websites. The main function is to analyze the users' emotions in micro-blogs and BBS forums, then push messages, but there are fewer fields for Web news hotspots. How to find a balance point of efficiency and accuracy is critical. We need to apply the existing clustering algorithm to the online clustering and algorithm improvement of news reports because this can enhance the efficient application of the algorithm. This can be a good search for hot news. In this way, it can provide a reliable way of information processing for the audience of the vast majority of the news, and it can also better accept the daily information.

References

[1] Han, X. H., Quan, L., Xiong, X. Y., et al. (2017). A novel data clustering algorithm based on modified gravitational search algorithm. *Engineering Applications of Artificial Intelligence, 61*(C), 1-7.

[2] Raza, M. Q., and Khosravi, A. (2015). A review on artificial intelligence-based load demand forecasting techniques for smart grid and buildings. *Renewable* and *Sustainable Energy Reviews, 50*, 1352-1372.

[3] Fong, S., Deb, S., and Yang, X. S. (2015). A heuristic optimization method inspired by wolf-preying behavior. *Neural Computing & Applications, 26*(7), 1725-1738.

[4] Myers, C. W. (2015). A New Genus and New Tribe for Enicognathus melanauchen Jan, 1863, a Neglected South American Snake (Colubridae: Xenodontinae), with Taxonomic Notes on Some Dipsadinae. *American Museum Novitates, 3715*(May 2011), 1-33.

[5] Inbarani, H. H., Bagyamathi, M., and Azar, A. T. (2015). A novel hybrid feature selection method based on rough set and improved harmony search. *Neural Computing* and *Applications, 26*(8), 1859-1880.

[6] Qu, X., Jain, A., Rajput, N. N., et al. (2015). The Electrolyte Genome project: A big data approach in battery materials discovery. *Computational Materials Science, 103*, 56-67.

[7] Hillman, S. L., Finer, S., Smart, M. C., et al. (2015). Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates. *Epigenetics Official Journal of the Dna Methylation Society, 10*(1), 50-61.

[8] Li, C., An, X., and Li, R. (2015). A chaos embedded GSA-SVM hybrid system for classification. *Neural Computing* and *Applications, 26*(3), 713-721.

[9] Thilagavathi, S., and Geetha, B. G. (2015). Energy aware swarm optimization with intercluster search for wireless sensor network. *The Scientific World Journal,* 2015 (3-30), 2015, 1-8.

[10] Agrawal, J., Agrawal, S., Singhai, A., et al. (2015). SET-PSO-based approach for mining positive and negative association rules. *Knowledge & Information Systems, 45* (2), 453-471.