A Hybrid Mechanism for Auto Text Categorization in Web Documents

Manas Kumar Yogi[1], Ch Manikanta Kalyan[2], Dwarampudi Aiswarya[3]

Asst. Prof. Computer Science and Engineering Department Pragati Engineering College (Autonomous) Surampalem, A.P., India.

E-mail: manas.yogi@gmail.com

Abstract

Web personalization has become such a popular paradigm nowadays, that almost all e-commerce websites are including it in their websites. The main objective of web personalization is driven by grouping similar web pages. The text categorization principle becomes a challenge when daily users visit numerous pages. This paper develops a hybrid framework which categorizes the text extracted from a web document, by applying Neighbourhood Preserving Embedding algorithm and then Particle Swarm Optimization algorithm on the extracted text groups, resulting into a group of web documents which contain similar texts. The proposed mechanism relatively has a high performance which improves with time, and as the size of web documents increase, the particle swarm algorithm also evolves in its nature.

Keywords: Embedded, particle swarm optimization, dimensionality, text categorization

1. Introduction

With the unsteady improvement in record data on the web and the speedy addition of PC execution, recovering archives from massive report databases has gotten considerable attention over the latest several years by virtue of its wide applications in various fields. As the web is inescapable, it has become basic for the clients to get the required information on the web [1-2]. It prompts huge improvement of web research on the internet. Although web surfers predict that information would be needed by the clients based upon their inquiries, most web users are not content with the recovery of the resultant records and the speed of the web lists. Since the web dataset is scattered, users to a great extent disregard to recover more relevant records for the client requests despite they are open in structure. Or on the other hand

DOI: https://doi.org/10.36548/jscp.2022.4.006
Received: 18.11.2022, received in revised form: 23.12.2022, accepted: 10.01.2023, published: 20.01.2023

perhaps they are introducing unessential web encyclopedias. Lately, huge analysis has been used to work on the execution of substance based recovery report. In evident recovery report structures, the significant inputs given by the client is habitually limited; however the archive space is reliably and outstandingly of high dimensionality. High-layered report data generally prompts substandard recovery results in light of the impact of dimensionality. A standard way to deal with this issue is to use dimensionality reduction techniques. Dimensionality reduction could satisfactorily improve the impact of dimensionality, further developing the execution and computational viability of report recovery system. Thus, it is appealing to initially task the high-layered record into a lower-layered subspace, in which the semantic design of the report space turns out to be clear; by this, the traditional recovery report calculations can by then be applied in the reduced lower-layered report space.

Latent Semantic Indexing (LSI) and Latent Discriminant Analysis (LDA) are two commendable devices extensively used in the report recovery system for dimensionality reduction and component extraction. LSI relies upon lower rank gauge of the term-record network from the Singular Value Decomposition (SVD), and it targets finding the best subspace gauge to the main archive space in regards to restricting the overall miscalculation. In spite of the way that LSI has been effectively applied to various archive information collection tasks, it doesn't consider class structure in the report data and is computationally enormous. Also, the choice of the ideal reduction estimation is difficult to speculatively decide. LDA is a managed dimensionality reduction calculation which intends to save discriminative information between data of different classes. LDA encodes one-sided information, by searching for headings that improve and meanwhile restrict the in-class disperse. However, one disadvantage of LDA is that it encounters the single sample size or under inspected issue.

In development, both LSI and LDA suitably determine the overall Euclidean construction. They disregard to find the fundamental complex construction concealed in the enveloping record space [3]. The Neighbourhood Preserving Embedding (NPE) is an actually proposed complex learning calculation for dimensionality reduction. NPE hopes to save the local complex construction; it can have serious isolating power. The NPE calculation has been successfully applied to various model selection issues. Eventually, the use of the calculation for web record recovery is at this points at an investigation zone where very few people have endeavoured to investigate. This work proposes an original programmed text order component considering NPE and Particle Swarm Optimization (PSO) calculation. The high-layered record data begins to extend into lower layered highlight space with NPE, and the PSO calculation is then applied to recover material reports in the reduced lower-dimensionality archive including space. Numerous analyses on a real educational assortment define the accuracy of the proposed calculation.

1.1 Applications of Text Categorization

- Enable users to an easy and efficient search within a website.
- Improvisations of browsing applications of web users by categorizing their search activity.
- Content curators in e-commerce companies save lot of time during item categorizing.
- Propulsion of emergency response systems due to categorization of discrete panic talks on social media.
- Improving the performance of search engine optimizer used by multiple popular search engines.

1.2 Need for Text Categorization

- Most data to be manipulated in modern day data processing systems are unstructured in nature, and dealing with it consumes the time of useful purposes.
- To navigate through huge amount of data in web space, text categorization plays important role in increasing the easiness of navigation.
- CRM tools of various top business companies make use of text categorization to add business value to the data available in their enterprise.
- Business organizations can make cost effective data management decisions by using the text categorization principles.

2. Related Works

2.1 Neighborhood Preserving Embedding

NPE is an as of late evolved dimensionality reduction component under the extent of complex learning and example acknowledgment. Dissimilar to LSI and LDA which attempt to decide only the worldwide Euclidean construction, the principal objective of NPE is to figure out the neighborhood complex design. By the utilization of NPE, the high-layered web records can be planned into a lower-layered highlight space in which the web reports with a similar class are in nearness to one another. Thus, NPE is supposed to have more specific power than LSI and LDA.

A sample of web pages with data {wd_1,wd_2,…wd_n $\subset R^p$ and M=[m_1,m_2,…,m_n] is considered. NPE determines a sequential change represented as a=R^{pxd}, which links each array, m_i(i=1,2,…n) in the p-dimensional space to a array y_i in the lower d-dimensional space by $y_i=a^Twd_i$ in such a way that y_i denotes wd_i with respect to some optimal heuristics, where Y=[y_1,y_2,...,y_n].

Given a weight matrix W, NPE can be derived by solving for a maximum value for the below given problem:

$$a_{opt}=argmax(a^TMM^Ta) \tag{1}$$

$$\text{subjected to, } YY^T=1 \tag{2}$$

$$\text{where, } M=(I-W)^T(I-W) \tag{3}$$

W denotes the weight matrix with Wij. The weights Wij on the edges can be computed by maximizing the below objective function:

$$\text{Max} \sum ||wdi - \sum \text{Wijwdj}||^2 \tag{4}$$

which is subjected to the limitation as shown below,

$$\sum_j Wij = 1, \text{ where i=1,2,…,n} \tag{5}$$

Note that the row arrays of X are sometimes linearly dependent, thus the matrix XX^T is singular.

2.2 Particle Swarm Optimization

PSO is a transformative calculation strategy, which is completely not the same as other developmental methodologies. PSO calculations don't include the sifting activities like hybrid and transformation, and the individuals from the entire populace are in contact using this technique. With regards to deciding an ideal or close ideal answer for the issue, PSO refreshes the ongoing age of particles (every particle is considered as a candidate answer for the issue) utilizing the information in regards to the best arrangement accomplished by individual particle and the whole populace. Every particle has a bunch of highlights which are current speed, current position, the best position recognized by the particle until this point and, the best position identified by the particle and its neighbors up to this point. Every particle begins by randomly instated speeds and positions. The PSO objective is to divide information between each of the particle of a populace. In PSO calculations, search is performed by using a populace of particles, in contrast to individuals [7]. Swarm optimization is applied as another space for effective web data recovery. Analysts have found that swarm optimization can be applied to take care of the optimization issues of auto text arrangement process.

PSO depends on the trading of information among people which are called particles of the populace, namely swarm. There are two variations of the PSO calculation which were upheld, one with a worldwide neighborhood, and other one with a nearby neighborhood. In the worldwide region, each particle moves towards its best past position and towards the best particle in the whole swarm, called gBest model. Of course, according to the nearby variety, called lbest model, each particle moves towards its best past position and towards the best particle in its bound neighborhood. Every particle moreover changes its own position reliant upon its previous experience and towards the best past position got in the swarm [8-9]. Recalling its best own position develops the particles experience deducing a nearby pursuit close by worldwide request emerging out of the adjoining experience or the experience of the whole swarm.

Table 1. Existing tools for text categorization

Reference No.	Attainment of the Model	Relevant Research issues
[4]	Weka-user friendly, applied for both supervised and unsupervised data	For large datasets, classification error is nearly 20%
[5]	SVM algorithm- High accuracy during label categorization	As length of labels increases, accuracy reduces
[6]	Fuzzy logic -based categorization of text- Corpus of sentences with specific sentiments categorized with high accuracy	Probability of misclassification out of domain words or labels is high.

3. Proposed Mechanism

In this research work, the following steps are taken to address the research issues mentioned in Table 1.

- Classification error is reduced to less than 15% for large datasets.

- Applying PSO algorithm to handle the dynamic evolution of labels given as part of user query during searching.
- Applying NPE mechanism which is comparatively less sensitive to Principal Component Analysis (PCA) technique. PCA technique is used majorly in all the popular text categorization methods.

This section presents a productive message classification calculation by utilizing the Vector Space Model (VSM) in which records are addressed as vectors. The text arrangement calculation has the accompanying two stages. In the first place, the high layered report space is planned with a lower layered highlight space; the PSO calculation is then applied to look through a bunch of records which best match the client's necessity by investigating various districts of the record space simultaneously [10]. Text categorization is at last an undertaking of finding fitting information from a more critical assortment of unstructured information. Text categorization measure runs out of sight, utilizes a gigantic records vault and the client question demonstrating client needs as info and recovers the most important archives at the top as yield. A record can be an organized information, text, video, image, sound, musical notes, DNA order, and so on. A report is typically portrayed by a bunch of catchphrases or terms contained in it. The client inquiries and reports should be spoken according to a model. Vector space model is broadly utilized, wherein vectors of term loads portray the two reports and inquiries. Vector space encases all the terms that the framework goes over and is built during ordering measure. Term weight assigns the criticalness of the term in the inquiry or in the record. For the given question, similarity values for the reports are processed utilizing a similarity measure. A few procedures are utilized to rank between records dependent on the processed similarity [11]. The highest level reports are considered pertinent to the inquiry and introduced as yield.

3.1 The text categorization algorithm involved in this mechanism is as shown below:

1) Conversion of the high-dimensionality document to lower dimensionality document using NPE algorithm.

2) Swarm structure initialization: The individual particle (lower dimension web page) is represented using document vector space model. Each particle is of the below indicated form:

$$Z_u(z_{u1},z_{u2},..z_{ur})$$

Where, r is total number of stemmed terms automatically pulled out from the lower dimension web documents, and z_{ui} is the weight of the i^{th} term in Z_u.

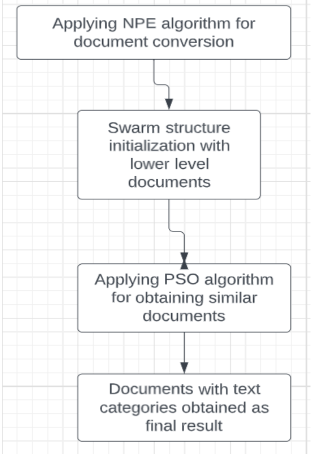
Applying NPE algorithm for document conversion
Swarm structure initialization with lower level documents
Applying PSO algorithm for obtaining similar documents
Documents with text categories obtained as final result

Figure 1. Flowchart of the Proposed Hybrid Framework

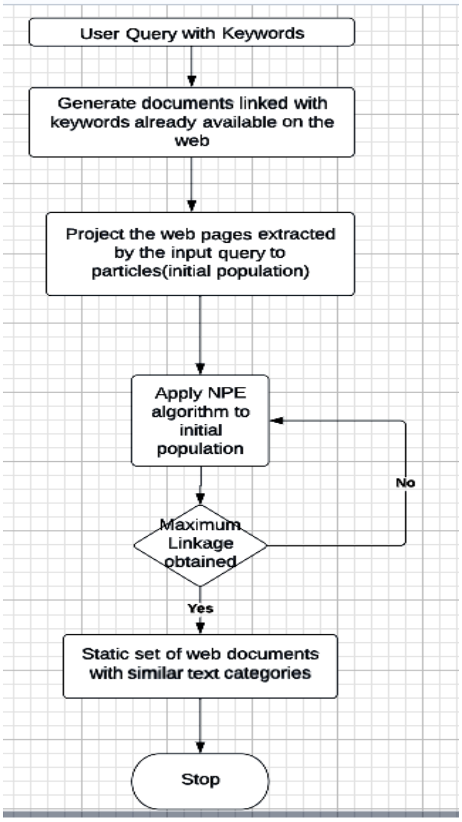
User Query with Keywords
Generate documents linked with keywords already available on the web
Project the web pages extracted by the input query to particles(initial population)
Apply NPE algorithm to initial population
No
Maximum Linkage obtained
Yes
Static set of web documents with similar text categories
Stop

Figure 2. Process flow of PSO algorithm in the proposed hybrid framework

The flow of process in the PSO algorithm application part of the system is presented below:

1) The framework takes user input as a query with some key terms.

2) The keywords are linked with a set of keywords already available in the web.

3) The web pages extracted by the input query to particles (initial population) are projected.

4) Consequently, the particles will undergo processing by the NPE algorithm which results into generation of a new population.

5) Step 3 is repeated when the maximum linkage value is obtained. The query returns a static set of web documents. No further changes are observed.

6) Stop the process.

In step 4, the NPE algorithm takes care of document vector mapping by converting the high dimensional web documents with low level dimension web documents. The low level dimension web document contains the terms matching the terms in the user query. The more common terms are categorized together by high level of coupling between the two web documents.

4. Experimental Results

The experiments were conducted to observe the comparative performance of various popular web retrieval techniques with the proposed framework (1) RF (Relevance Feedback), (2) NPE-ACO(Neighbourhood Preserving Embedding-Ant Colony Optimization) and (3) ACO (Ant Colony Optimization) method. As users need to give input unequivocally in significance analysis, users don't have to put effort in this proposed method, since it will straight forwardly store changes in client's inquiry profile. Reuters-21578 Text Categorization Collection Dataset is used to apply the proposed mechanism. Python is used as the programming langauge to implement the proposed framework on Windows 10 operating system with 8GB RAM.

Table 2. Comparative Analysis of Popular methods

Technique Used	Normalization scope over 2 weeks	Normalization scope over 4 weeks	Normalization scope over 6 weeks	Normalization scope over 8 weeks
Relevance Feedback	0.075	0.081	0.088	0.094
ACO	0.062	0.074	0.085	0.089
NPE-ACO	0.064	0.078	0.084	0.086
NPE-PSO	0.061	0.073	0.080	0.078

In the experiments, the biggest 20 classifications which have the largest number of records from other famous techniques were chosen. The presentation of the 4 strategies were plotted for a duration of nearly 4 weeks against the level of web documents recovered. It was discovered that over a time of 2 weeks, three strategies were performed with practically equivalent capacity. But in following 2 weeks, the proposed Neighbourhood Preserving Embedding-Particle swarm optimization (NPE-PSO) technique edged past the other two techniques by around 2 to 3%. By the end of 6 weeks, the categorization rate of the proposed framework was significantly higher than the other 3 popular procedures.

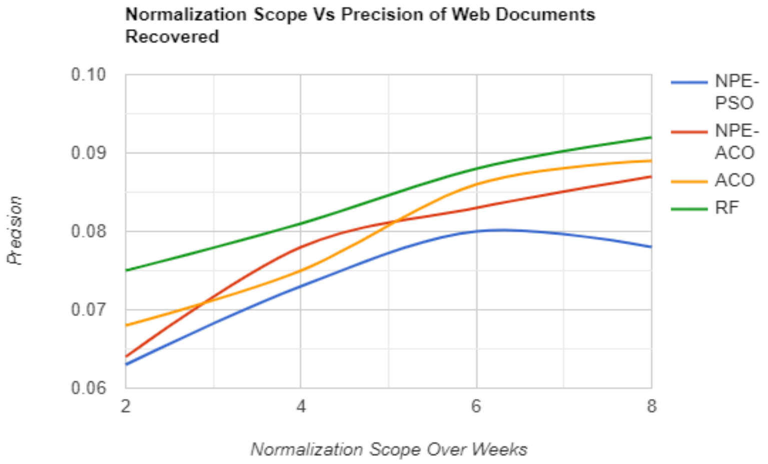

Figure 3. Plot of Scope Versus Precision

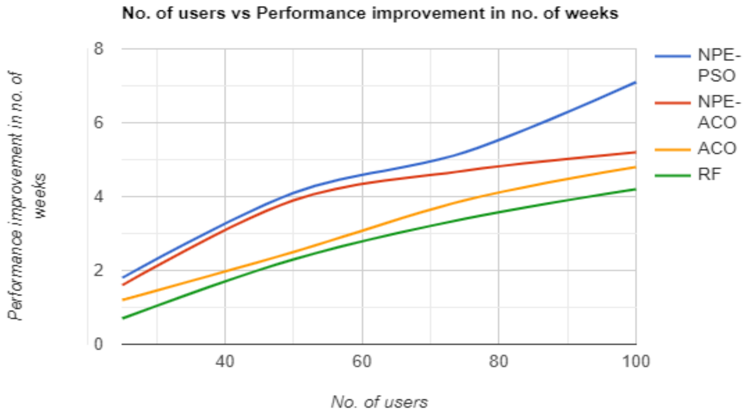

Figure 4. Plot of Number of Users Versus Performance over time

Table 3. Comparative Analysis of Popular methods

Technique Used	Performance improvement with 40 users	Performance improvement with 60 users	Performance improvement with 80 users	Performance improvement with 100 users
Relevance Feedback	1.5%	2.8%	3.8%	4.1%
ACO	1.7%	3%	4%	4.4%
NPE-ACO	3%	4.1%	4.8%	5.1%
NPE-PSO	3.1%	4.3%	5.2%	7.5%

Table 4. Experimental Outcomes after Text categorization

Document Name	Number of Documents Present	No. of Labels	No. of Categories obtained after application of NPE-PSO method
reut2-000.sgm	1000	22	202
reut2-001.sgm	1000	14	178
reut2-002.sgm	1000	28	145
reut2-003.sgm	1000	21	139
reut2-004.sgm	1000	19	481
reut2-006.sgm	1000	33	312
reut2-007.sgm	1000	36	159
reut2-008.sgm	1000	20	155
reut2-009.sgm	1000	17	287
reut2-0010.sgm	1000	24	227

5. Conclusion

To expand the current research work done in the field of auto text categorization, the Neighbourhood Preserving Embedding-Particle swarm optimization (NPE-PSO) technique has been proposed in this work. From the experimental results, the proposed NPE-PSO framework is found to perform a rate which is almost 2.2% more than the other three well-known strategies for auto text categorization. In future, the observational period could be extended to a longer span and can be checked if the rate increment in retrieval is expanding or staying stable. The primary favorable basis of utilizing this proposed mechanism is that its performance doesn't degrade regardless of whether the search space is increased. Regardless of whether search space is expanding or then again diminishing, the retrieval rate doesn't drop for the proposed strategy. In future, this work would be extended by combining this proposed strategy with user history based search profile so as to further decrease the retrieval time.

References

[1] Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, pp. 39–43 (1995)

[2] Kennedy, J.: The particle swarm:social adaptation of knowledge. In: Proceedings of 1997 IEEE International Conference on Evolutionary Computation, Indianapolis, pp. 303–308 (1997)

[3] Salton, G., Wang, A., Yang, C.S.: A vector space model for information retrieval. Journal of the American Society for Information Science 18, 613–620 (1975)

[4] Donatella Merlini, Martina Rossini,Text categorization with WEKA: A survey,Machine Learning with Applications,Volume 4,2021,100033,ISSN 2666-8270,https://doi.org/10.1016/j.mlwa.2021.100033.

[5] Dhar, Ankita, et al. "Text categorization: past and present." Artificial Intelligence Review 54.4 (2021): 3007-3054.

[6] Sathe JB, Mali MP (2017) A hybrid sentiment classification method using neural network and fuzzy logic. In: Proceedings of IEEE international conference on intelligent systems and control, pp 93–96.

[7] Robertson SE, Jones KS (1976) Relevance weighting of search terms. J Am Soc Inf Sci 27(3):129–146

[8] Robertson SE, Walker S, Beaulieu M, Gatford M, Payne A (1995) Okapi at trec-4. In: Proceedings of the 4th Text Retrieval Conference, pp 73–97.

[9] Rocchio JJ (1971) Relevance feedback in information retrieval.The SMART Retrieval System - Experiments in Automatic Document Processing, pp 313–323

[10] Salehi S, Selamat A, Mashinchi MR, Fujita H (2015) The synergistic combination of particle swarm optimization and fuzzy sets to design granular classifier. Knowl-Based Syst 76:200–218

[11] Saraiva PC, Cavalcanti JM, de Moura ES, Goncalves M. A., Torres RDS (2016) A multimodal query expansion based on genetic programming for visually-oriented e-commerce applications. InfProcess Manag 52(5):783–800