Traditional techniques for web content classification

T.SREENIVASULU
Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Chennai, Tamilnadu, India
sreesree82@gmail.com

Dr.R.JAYAKARTHIK
Assistant Professor
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Chennai, Tamilnadu, India
drrjayakarthik@gmail.com

Abstract

In today's advanced world of internet era, WWW became powerful for saving and retrieving information. Due to variety and unorganized nature of data on WWW, searching information is becoming awkward & time taking task. Web mining aroused as a solution for the above problem. Web content mining is a subset of web mining. Data Mining has gained familiarity in various fields and classification is become more important. In order to classify data or content, various techniques are available and applied. Each technique has pros and cons. This paper discuss about overview of the traditional web content techniques and conclude with algorithms and their achievements used for web content classification.

Keywords: *Web Mining, Web content mining, Web content classification techniques and algorithms, Data mining.*

1. Introduction:

Web mining is the Data Mining methodology that consequently finds or thinks the information from web reports. It is the extraction of entrancing and potentially important models and certain information from antiquated rarities or activity related to the overall web. Web mining is the incorporation of information gathered by standard data mining systems and methodology with information aggregated over the World Wide Web. Web-mining is a multi-disciplinary effort that attracts strategies from fields like development recovery, estimations, AI, natural language, and others.

The web mining turns into the difficult undertaking because of the heterogeneity and absence of structure in web assets. On account of these circumstances, the user presently suffocating in data and confronting data over-burden [1]. The greater part of the web user could experience the accompanying issues, while collaboration with the web;

A. Information Finding:

At the point when a user needs to discover explicit data in the web, they input a basic watchword question. The inquiry reaction will be the rundown of pages positioned relies upon their closeness to the question. However, the present inquiry devices have a few issues, for example, Low exactness (because of the insignificance of list items) and Low review (powerlessness to record all the data accessible).

B. Formation of New Knowledge:

This issue is an information activated procedure though the past is a question activated procedure. Here the user needs to remove conceivably valuable data from an assortment of accessible substance.

C. Data Personalization or Customizing Data:

This is related with the sort and introduction of data, as almost certainly, individuals vary in the substance and introductions they like while communicating.

D. Understanding User Preferences:

This work with the issue of experiencing the requirements of users. This incorporates personal preferences of individual user, website plan and the executives, redoing client data and so forth. The web gets uproarious on the off chance that it contains different sorts of data. The web mining procedures can be utilized to understand those issues.

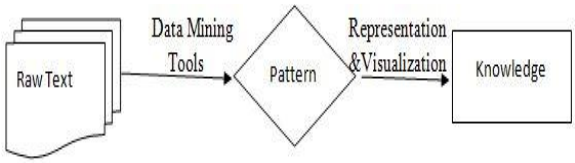

Fig. 1. Web Mining Process

2. Web Mining Categorization:

Web Mining is categorically divided in to 3 types based on the format of the web data: Content, Structure and Usage mining.

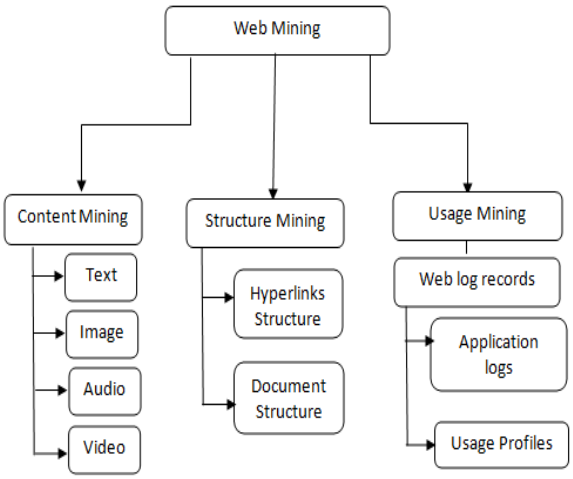

Fig. 2. Web Mining Categorization

Content Mining:

The viewable information on the website pages or any sort of data which incorporates content, sound, video, pictures, HTML, XML is known as the substance. To get these kinds of information from various website pages goes under Web Content Mining. Web Content Mining includes exhuming organized information, semi organized information or non-organized information

Structured Mining:

It is a device polished to find the connection between web pages related with data. The fundamental goal of web structure mining is to take out the recently known connections between the web pages. It essentially utilizes the diagram hypothesis with different nodes and the association to all the nodes. In the field of business or E-Commerce, a group of clients i.e. clusters can be made for looking through comparative kind of information on the web which brings about progress in a various organizations productively and increment in the sales.

Usage Mining:

Usage Mining is the case of obtaining any sort of data from server logs [2]. It is the way toward examining the interest of the clients on the web i.e. in what sort of information they are intrigued for. For example, few users are keen on content sort information or some web users are keen on sound, video or pictures. Web Usage Mining assists to find out user behaviour. Web usage mining makes it easier to users by getting the distinctive kind of recommendations for which they are searching for .E.g. Online Shopping for a specific item, Property Search and so forth.

3. Web Content Mining Classification Techniques:

3.1 Hybrid of particle swarm optimization algorithm:

The huge quantities of a book archive can be gathered with the assistance of text bunching procedure. The impact of the text grouping decrease efficiency by changing the size of the Document. In the event that the content of text comprises of the uncommon data, the grouping calculation execution is diminished on the record and furthermore the calculation time diminished. The fundamental innovation of unaided learning known as highlight determination which is applied for choosing

another arrangement of important content capacity to decrease the calculation time and furthermore improve the exhibition of the content groups. The hereditary administrator is used for the upgrading a swarm of particles by a calculation which utilized right now, (Abualigah L.M. also, Khader, A.T., 2017) in which the hereditary administrator reasonable to character choice issues. K-implies grouping was utilized for getting the legitimacy of the capacity of the subsets. The pre-preparing steps are used in the initial phases in which the advanced test record are introduced then which are changed into the content archive and afterward the numerical style was shaped from the content report in a last which are changed into the numerical grid. Uninformative content highlights are dispensed with in the second steps by the use of a mixture of the PSO to upgrade the content grouping. Right now, administrator of the hereditary calculation joined with the PSO calculations for modifying the arrangement of PSO. Bunching calculation required only a few backings to give precise clusters is the benefit of this method [3].

3.2 Ant colony optimization (ACO) algorithm:

The ant's natural behaviour persuaded the Ant Colony optimization algorithm (ACO) in which the ant natural behaviour is utilized in the deciding of the ideal path from nets to the origin of substance. The ACO strategy used for comprehending the detailing of enhancement of the grouping issue (Forsati, R. et.al. 2015). The clustering of information was directed by multi-ant colony approach which include a sovereign ant specialist and some equal and independent ant colonies. Different form of likelihood change work and different sorts of ants moving velocity are taken at the each ant colony process for delivering the various outputs of clustering. The literature have an alternate type of hybrid algorithms which depend on the ACO strategy. Clustering of web session was led by Ant Clust algorithm. In this work report recovery done by the assistance of ACO based grouping [4].

In this approach, dropping or assembling of proposed archive vectors done by the mapping of the walking around ants in which gathering and dropping are led with different probabilities. Zhang et.al present another work in which the arbitrary movement of ant was estimated in the space of solution which additionally presents join with a moderate rate. In this, AFTC was known as the quicker document grouping. The ants organize the pheromone to limit the undesirable arbitrary movement that gives the control in each progression of the procedure to the ants for moving the particular direction through high pheromone focus. The first class was held by the system that known as the chore of algorithm in which a specific number of the important result was holding at the every one of reiteration in the method of an algorithm for entering them into the following pattern of the algorithm. This procedure are used for improving the exhibition of the algorithm

3.3 Genetic Algorithm:

The Genetic Algorithms are known as the pursuit techniques which work dependent on the regular theory of advancement. In the genetic algorithm, the limited length of the string of certain cardinality was acquired by directing encode procedure on the choice factors of web crawler. These strings speaking to the applicant answer for the issue are alluded to as chromosomes. The qualities are alluded by the letter sets of the string in which the populace was known as chromosome and the estimations of qualities are named as the alleles. The presentation of the genetic algorithm was upgraded by the GA which knows as the client characterized parameters which used the size of the populace [5]. The untimely combination and an imperfect results are shaped by the small populace size while the lot of computational impacts would occur by the enormous populace size. The size of the high populace and the low populace was not picked in the process for disposed of the elements, for example, high computational overhead and untimely intermingling (Chawla, S., 2016).

3.4 Harmony search based Pareto optimization (HSPO) algorithm:

The HSPO procedure is named as starting Pareto ideal solutions which are recommended to make Pareto ideal answers for the deterministic MMOP issues (Guo, Z.X. et.al. 2015). This procedure incorporates 7 methodologies. An amicability search process was incorporated from the non-commanded arranging method with the assistance of the HSPO procedure in which the concordance procedure coordinated to make Pareto ideal answers for the deterministic MMOP issue [6].

1) Parameter introduction
2) Harmony memory introduction
3) Performance assessment of the newly created harmony
4) Improvisation
5) Harmony arranging utilizing a non-overwhelmed arranging strategy
6) Updating of agreement memory
7) Termination criteria checking

The procedures are utilized in the creation that for the most part relies upon the starting time which additionally compares to the appearances time of the procedure, and furthermore the hour of procedure for completing and furthermore the significance of the procedure consistently relies upon the requests. Some creation division has anticipating and lingered for the appearance of requests to deliver the items. The necessary results of early coming requests created first. In the event that a branch of the organization has a lot of request to work, the most elevated procedure need work continues first. The beneath referenced principles are followed right now for choosing the procedure need of each request and each request gathering.

1) The right on time due date arranges in the arranged gathering list has higher need to continue.
2) If products request bunches have the equivalent due date, the less creation remaining task at hand required request bunches item will delivered with a higher need.
3) The request with the bigger number of procedures is delivered with a higher need in a request gathering.
4) The request with less task at hand is delivered with a higher need in the event that that numerous requests have a similar number of creation forms in a request gathering.

3.5 Scale-Free Binary Particle Swarm Optimization:

A completely associated system shaped by a molecule for connecting each other at the predefined interim in the traditional PSO which is not the case in real time systems. A scale -free BA model introduced and clarified by, Gupta, S.L. et.al in 2019 in which human-made systems and genuine systems are totally associated by scale free structure rather than homogeneous standard system and neither totally associated. The power law conveyance is appeared by the level of hubs in the network, for example, World Wide Web, reference systems, programming building, online informal communities and Internet. The low degree hubs which known as low compelling and exceptionally persuasive hubs are firmly associated by the one a couple of hubs. The effect of without scale topologies on the exhibition of PSO was examined and discovered a lot of viable than standard PSO. The fundamental standards of the without scale arrange works are "particular connection" and "development" [7].

3.6 Particle swarm optimization algorithm:

In present, the best system for the streamlining procedure known as swarm optimization (PSO) algorithm which at first settled by Eberhart and Kennedy (Younus ZS, et.al. 2015). On the off chance that an operator flying over the issue, the bird flocking rushing symbolizes particles are reflected. In any case, a molecule's area of multidimensional issue space implies an answer, however as particles move to another area, the arrangement is assessed by a wellness work that gives a quantitative estimation of the result’s utility [8].

2D bird flocking simulated by Eberhart and Kennedy to plan the PSO. The situation of every specialist was meant by x and y, while vx indicate the speed (the speed of the x-hub) and vy otherwise called the speed along the y-pivot i.e. position of a specialist changed

comparing to data and speed. The bird flocking upgrade process was used by lot of administrators to discover the best ideal value and its situation along the x and y-direction. Each and every administrator recognizes the best an incentive in the gathering (gbest) among pbest.

Besides, the PSO algorithm have five conscious the five basic parameters. While the activity point by point on particle swarm optimization is given beneath.

The algorithm steps include:

Stage 1: Initialization

Stage 2: Velocity refreshing

Stage 3: Position refreshing

Stage 4: Memory refreshing

Stage 5: Termination checking

4. Summary Evaluation

Table.1 Comparing Machine learning methods in web content classification

Technique	Advantages	Drawbacks
Hybrid of particle swarm optimization algorithm	Encourages the clustering algorithm to obtain accurate clusters	The documents size affects the text clustering by decreasing its performance
Ant colony optimization (ACO) algorithm.	Reduces the time required to classify new web pages sharply without loss of accuracy in classification	A slight variation on an edge allows the edge to be chosen.
Genetic algorithm.	To avoid both premature convergence and high computational overhead	A large population size would involve a lot of computational effort
Harmony search based Pareto optimization (HSPO) algorithm	The order group with an earlier due date is produced with a higher priority	One production department cannot perform more than one production order at a time
Scale-Free Binary Particle Swarm Optimization	Complex and Time consuming.	Over-fitting problem and have poor accuracy

Table.2 Performance comparison of PSO-based approaches across different datasets

References	Approach	Dataset	Purposes	Result
Gupta SL, (2019)	Scale-Free Particle Swarm Optimization (SF-PSO)	Six high-dimensional datasets	Increase the classification accuracy and to reduce the time complexity	Classification accuracy 70%
Abualigah LM, and Khader AT. (2017)	Hybrid of particle swarm optimization algorithm	Eight common text datasets	Partition a huge amount of text documents into groups.	-
Carneiro MG et.al. (2019)	Particle swarm optimization	Vector-based data set	Optimizing a quality function driven by the classification accuracy	-

5. Conclusion

This paper incorporates the strategies for categorization which are usually utilized to mine the data from the web, Each having its own favourable circumstances and drawbacks. The choice of the strategy relies upon the application. For the future work other grouping strategies can be considered to improve classification performance.

6. References

[1] P. Maes. Agents that reduce work and information overload. Communications of the ACM, 37(7):30–40, 1994.

[2] M onik a Yad a v , Pra d eep Mi tt a l , " Web Mining: An Introduction ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013 ISSN: 2277 128X.

[3] Abualigah LM, Khader AT, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering", The Journal of Supercomputing, vol. 73, no. 11, pp. 4773-4795, 2017.

[4] Forsati R, Keikha A, Shamsfard M, "An improved bee colony optimization algorithm with an application to document clustering", Neuro computing, vol. 159, pp. 9-26, 2015.

[5] Chawla S, "A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search", Applied Soft Computing, vol. 46, pp. 90-103, 2016.

[6] Guo ZX, Yang C, Wang W, Yang J, "Harmony search-based multi-objective optimization model for multi-site order planning with multiple uncertainties and learning effects"' Computers & Industrial Engineering, vol. 83, pp. 74-90, 2015.

[7] Gupta SL, Baghel AS, Iqbal A, "Big Data Classification Using Scale-Free Binary Particle Swarm Optimization", In Harmony Search and Nature Inspired Optimization Algorithms, pp. 1177-1187, 2019.

[8] Younus ZS, Mohamad D, Saba T, Alkawaz MH, Rehman A, Al-Rodhaan M, Al-Dhelaan A, "Content-based image retrieval using PSO and k-means clustering algorithm", Arabian Journal of Geosciences, vol. 8, no. 8, pp. 6211-6224, 2015.