International Journal of Database Theory and Application
Vol.8, No.1 (2015), pp.63-74
http://dx.doi.org/10.14257/ijdta.2015.8.1.08

Review: Biological Optimization Techniques in Webpage Classification

Shashank Dixit[1] and Dr. R. K. Gupta[2]

[1]Department of Computer Science & Engineering, Mits, Gwalior, India
[2]Professor, Department of Computer Science & Engineering, Mits, Gwalior, India
[1]shashankdixit54@gmail.com, [2]iiitm_rkg@rediffmail.com

Abstract

With the explosive growth of the data stored in various forms, need for innovative and effective technologies to find and use information and knowledge from a large variety of data sources which is continually increasing. Web information contains a lot of noise. Web Mining is the application of data mining techniques to discover classification of web data. It focuses on techniques that could predict the data's class while the user interacts with the web. The aim of this paper is to measure, propose and improve the use of advance web page classification techniques which is highly used in the advent of mining large web pages based data sets which allows data analysts to conduct more efficient execution of large scale web pages data searches. Thus in this paper researchers introduce an improved concept which may reduce the search space using classification techniques with optimization technique.

Keywords: *Web Mining, Classification, Data Mining Techniques, Optimization Techniques, Feature Extraction*

1. Introduction

Data mining is used to extract useful information or pattern from huge amount of data. Which is further used for classification and prediction task [5]. Data mining can be applied where data stored in tabular form, relational table, spread sheets, and text and web data. Web mining become important because data are increasing on the web day by day in large amount.

Some issues are present in web page classification existing techniques Optimizing the within classification variation is computationally challenging .The necessarily for users to specify, the number of classification in advance can be seen as a disadvantage. The numbers of web pages at various servers are alike thus classification the data according to the relevance requires a large amount of query system. Outlier at web page cannot be detected while large number of filter is used. They are unrecognizable.

Here outliers are redundant web pages and thus they have to be out clustered and sorted according to the relevance.

Web page classification is a task of text classification from webpage we have to find useful pattern or feature from web page to classify it. Web page can be classified by two types such as manual classification and automatic classification. Manual classification is a slow process it is time consuming process for large data set while automatic classification can easily handle large data set and it is faster and accurate because it is based on algorithms.

The problem of Classification is to decide from which class a given record belongs. There are two type of classification I.e. unsupervised and supervised .In unsupervised classification class label of training data is unknown i.e. from which class a given record belongs is not known while in supervised classification class label of training data is known there are two phase training phase, testing phase. In first phase set of training

ISSN: 2005-4270 IJDTA

record is used to train the classifier while in second phase set of document is used to test the classifier [5].

Classification plays a vital role in many information retrieval and management tasks. Web page classification can improve the searching quality of web search. Problem can be represented in different form based on classes, classification can be divided into two types *i.e.*, binary classification and multiclass classification, binary classification categorizes instances into exactly one of two classes, and multiclass classification deals with more than two classes. It depends on the number of classes that can be assigned to an instance, classification can be categorized into single class label classification and multi-label classification. In single class label classification, one and only one class label is to be assigned to each instance, while in multi class-label classification, more than one class can be `assigned to an instance.

While researchers talk about the various classification methods, it is very obvious that there are so many other optimization techniques are also available to increase the efficiency of the various available methods.

In this paper, the relationships among the techniques of data mining, web mining and optimization techniques are studied and used for web page classification.

The rest of the paper is organized as follows: Section 2 briefly introduces various data mining techniques. Section 3 briefly introduces the web data mining and the web classification process. Section 4 provides various optimizing techniques. Section 5 contains the comparison of different technique. Section 6 contains support vector machine and Section 7 role of support vector machine (svm) in web classification optimized by firefly and Section 8 provide conclusion and future work.

2. Data Mining Techniques

Data mining uses a relatively huge amount of computing power operating on a large set of data stored in repositories to determine regularities and connections between relevant data points [3]. To search large databases we can use the techniques called statistics, pattern recognition and machine learning are used to search large databases automatically. Another word for data mining is Knowledge-Discovery in Databases (KDD) [4, 5]. The data mining helps bank or any other organization to increase its ability to gain deeper understanding of the patterns previously unseen using current available reporting capabilities. Further, prediction from data mining allows the bank or any other organization an opportunity to act with customer drops out or top loan for resource allocation with confidence gained from knowing how to interact with a particular case [3].

3. Web Data Mining and Web Classification

Data mining is the study of data-driven techniques to discover patterns in large volumes of raw data. Web mining can be referred as the transformation of the data mining techniques to web data. Web mining is highly divided into various categories (see Figure 1) upon which our research is focused these are web page classification mining [16].

3.1. Web Structure Mining

Mining the content involves extracting the relevant information. Structure mining studies the structure and prototype. In other words, it is a process by which we discover the model of link structure of web pages. We catalog the links, generate the web pages generate the information such as the similarity and relations among them by taking the advantage of hyperlink topology. The goal of web structure mining is to generate structured summary about the website and web page. Page rank and hyperlink analysis also belongs to this category. It tries to discover the link structure of hyper links at inter document level [16, 17]. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that web structure mining has a relation with web content mining.

3.2. Web Content Mining

It is also known as text mining, is generally the second step in web data mining. Content mining is the scanning and mining of text, pictures and graphs of a web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query [16, 17]. With the massive amount of information that is available on the web, web content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

3.3. Web Usage Mining

Web usage mining is the analysis of the discovered patterns. In other words, it is the process from we can identify the browsing patterns by analyzing the navigational pattern of user. It concentrates on techniques that can be used to predict the user behavior while the user interacts with the web. It uses the secondary data on the web [16]. This activity involves the automatic discovery of user access patterns from one or more web servers. Through web usage mining we can find out what users are looking for on Internet [16].

The working of WUM has three steps – preprocessing of the data, pattern discovery and analysis of the patterns. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources not only discover quality patterns but also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented.

3.4. Web Page Classification Mining

Classification plays an important role in managing web directory. On the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific Web link analysis, to contextual advertising, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of Web search [18].

3.5. Web Page Classification Techniques

Web pages can be classified into the following categories:

(1) Manual classification

(2) Clustering approaches

(3) META tags based categorization

(4) Text content based categorization

(5) Link and content analysis

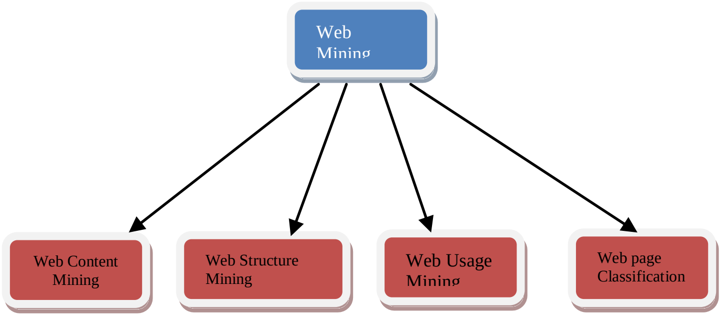

Figure 1. Classification of Web Mining

4. Optimization Techniques

This section serves as a quick review of nature-inspired algorithms. Readers who are interested in the full details about these algorithms and their integration mechanism are referred to the following inline citations.

4.1. Firefly Algorithm

It is a Meta heuristic algorithm, inspired by the flashing behavior of fireflies [6]. To attract other fireflies is the main aim of firefly's flash. Xin-She Yang formulated this firefly algorithm by assuming: 1.All fireflies are unisex, so that one firefly will be attracted to all other fireflies; 2. Brightness make them attractive accordingly, and for any two fireflies, the less brighter one will attract (and thus move) to the brighter one; here, distance increases makes decreases of brightness; 3.If there are no fireflies brighter than a given firefly, it will move randomly .So objective function must have brightness component. Recent studies show that FA is particularly suitable for nonlinear multimodal problems.

4.2. Cuckoo Search

It is an optimization algorithm developed by Xin-She Yang and Suash Deb in 2009 [7]. It was inspired by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of other host birds (of other species). Some host birds can engage direct conflict with the intruding cuckoos. For example, if a host bird discovers the eggs are not their own, it will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. Some cuckoo species have evolved in such a way that female parasitic cuckoos are often very specialized in the mimicry in colors and pattern of the eggs of a few chosen host species. Out of so many optimization techniques this CS idealized such breeding behavior works at most of the places or problems. CS uses the following representations: Each solution is presented by a egg in a nest, and so each new solution is presented by a cuckoo egg. Main purpose of this is to replace a not-so-good solution in the nests to use the latest and most probably better than other solutions (cuckoos) to. For the simplicity each egg is in each nest. In many applications, cuckoo search can outperform other algorithms such as particle swarm optimization and ant colony optimization. The algorithm can be extended to more complicated cases in which each nest has multiple eggs representing a set of solutions. Their invention “Novel 'Cuckoo Search Algorithm' Beats Particle Swarm Optimization” was recently reported at ScientificComputing.com [8].

4.3. Bat Algorithm

Bat-inspired algorithm is a Meta heuristic search optimization developed by Xin She Yang in 2010 [9]. This bat algorithm is based on the echolocation behavior of micro bats with varying pulse emission and loudness. The idealization of echolocation can be summarized as follows: Each virtual bat flies randomly with a velocity v_i at position (solution) x_i with a varying frequency or wavelength and loudness A_i at i_{th} step.

4.4. ABC Algorithm

Artificial Bee Colony (ABC) algorithm is a problem solving method, developed based on behaviors of honey bee colony, foraging and sharing the information with other colony members in the hive, to be able to exploit richest food sources in shortest possible time [10, 19, 20, 21].

A possible solution in the problem is represented by a position of a food source in nature. The nectar amount in that food source represents the fitness value of that solution.

There are 3 types of bees in ABC algorithm.

4.4.1. Employed bees, Ne (number of solutions in ABC algorithm in one iteration step)

4.4.2. Onlooker bees, No

4.4.3. Scout bees

4.4.1. Employed Bees

Every employed bee works on a food source. Position of this food source symbolizes a possible solution. Employed bee calculates the fitness of this solution and saves the position information in its memory. Number of employed bees in the hive “Ne”, represents the number of solutions in ABC algorithm in one iteration step.

4.4.2. Onlooker Bees

Those bees that are waiting in the hive Receive information about the position of food sources from employed bees. Each onlooker bee selects a food source to exploit, depending on the nectar amount. This selection is done by modeling the nectar amount as a probability (Pq, probability of q^{th} employed bee being selected by onlookers). The more nectar amount is exist, the higher probability that employed bee is selected. After this selection, each onlooker bee searches new position near the selected food source (employed bee location).

4.4.3. Scout Bees

A bee who is foraging new food sources without any information is called scout bee. They randomly search whole environment. A scout bee becomes an employed bee when it starts to work on a food source.

4.5. PSO Algorithm

PSO was originally developed by Eberhart and Kennedy in 1995 [22] is a population based global optimization technique, and it takes inspiration from social behavior of a birds flocking. In the PSO algorithm, the birds in a flock are shown as particles in n-dimension. Best fitness value of particle at a location in the n-dimensional problem space represents one solution for the problem. When a particle updates it’s position, another problem solution is generated and then new solution is evaluated by fitness function and the process is repeated until a stopping criteria is met.

Each particle is initialized with initial position and initial velocity in n-dimension space. Each particle represents solution for problem and each particle’s fitness is evaluated using simple formula. We have to define boundary for communication grouping between the particles then the best particle with best fitness value is selected as a local best solution it is represented as pbest. Then it is compared with other group’s pbest if it’s value is better than all then it is selected as the best feature it represent the best solution as global best gbest.

4.6. Intelligent Water Drops (IWDs)

It is a nature inspired technique developed by Shah-Hosseini in 2007 [23].In nature, flowing water drops are observed in rivers how they starts and how they find path from source to destination? The path that natural river takes from the action and reaction between water drops and riverbeds. Assume an imaginary natural water drop is going to flow from one point of a river to the next point in the front. Intelligent water drop have two parameters to maintain static and dynamic. Static parameters remain same for the lifetime of IWD while dynamic parameter changes after each iteration.

IWD is a graphical representation in which each node is initialized with IWD drops. Each IWD starts with initial soil and initial velocity and then it start removing soil by water drop with initial velocity. Each IWD maintains it’s memory and visited node list .if visited node is not found then it is added to IWD node list. Probability based evaluation is used to add new node in list.IWD prefers the path which having minimum soil then path having high soil. Here in last amount of soil is proportional to time taken from source to destination.

Based on the aforementioned statements, an intelligent water drop (IWD) has been suggested [23], which possesses a few important properties of a natural water drop which help to calculate.

This Intelligent Water Drop has two important properties:

(1) The soil carried by IWD, denoted by soil (IWD).

(2) The velocity that it posses, denoted by velocity (IWD).

4.7. Ant colony Optimization (ACO)

Ant colony optimization (ACO) is among the most successful swarm based algorithms proposed by Dorigo & Di Caro in1992.it takes inspiration from ant colony and behavior of ant for finding food by searching the shortest path. They travel from source to destination (food) after finding they return to their colony by laying down a substance called pheromone it is an attractive attribute and other ant follows the pheromone trail which has highest amount of pheromone. After some time the evaporation of pheromone starts this reduces the attractiveness of path. So that ant searches for the new path.

Ant miner algorithm is developed for web page classification. there are many version available for ant miner such as ant miner, ant miner 2,ant miner + *etc.*, in ant miner algorithm we have to represent attribute in graphical format and find classification rule which has two part if <antecedent> then <consequent> .antecedent is combination of categorical attribute and consequent represent class. some steps of the algorithm is as follows: Initialization of pheromone values and then Rule production is done after this Heuristic function is used to calculate the probability of path selection then Rule pruning is done to improve accuracy of rule and in last pheromone value is updated

This process is repeated until the best rules or features have been found from web page.

5. Comparison of Different Optimization Techniques

In this section we are representing comparative study of different optimization techniques including scalability, method based on, important parameter, advantage and disadvantage.

S. no.	Technique name	Scalability	Based method	Important parameter	Advantage	Disadvantage
1	**Firefly algorithm**	Local Search	Population Based Optimization	Light intensity, light absorption coefficient, population of firefly. Number of iteration	Automatic subdivision of local optima and global optima, less number of parameter, provide better convergence.	It sometimes stuck in local optima,
2	**Cuckoo Search**	Global Search	Population Based Optimization	Number of nest, number of eggs, fitness value, number of iteration	It balance efficiently local and global optima, uses less number of parameter. less complex	Number of iteration is so large to finding optimal solution, not so efficient
3	**Bat Algorithms**	Local Search	Metaheuristic Algorithms	velocity, position ,frequency, wavelength, loudness, pulse emission, number of bat	Automatic control of range and speed using frequency, it can be applied on various mathematical problem, very accurate and efficient	We have to limit the wavelength, challenge to solve problem of higher dimension.
4	**Artificial bee colony(ABC)**	Random Solution	Numeric Optimization	Colony size, worker and non worker bee, nectar amount	better convergence at local minima as well global optima, less control parameter	It is not suitable for large and complex problem.
5	**Particle swarm optimization (PSO)**	Multi- agent parallel global-search technique	Mathematical Modelling	Number of particle, particle position and velocity, particle range ,inertia weight, number of iteration	Centralized control to guide the particle, large number of particle easily handle,	huge number of parameter, no automatic subdivision of local and global optima we have to define boundary

6	**Intelligent water drop (IWD)**	Global-Search Technique	Mathematical Modelling	Velocity, soil ,number of iteration, static and dynamic parameter, water drop number's	flexible in dynamic environment ,higher accuracy and feasibility	large number of parameter then FA, only suitable for web services
7	**Ant colony optimization** **(ACO**)	Local search	Population based optimization	Number of ant, pheromone value, number of iteration, evaporation rate	Better then GA, neural network ,can be used in dynamic environment, retain memory of entire colony	It perform poor for large instance problem, slower convergen ce then other method, velocity is irrelevant to algorithm

6. Support Vector Machine

The SVM developed by Vapnik [15]. The support vector machine [SVM] is a training algorithm. It trains the classifier to predict the class of the new sample. SVM is based on the concept of decision planes that defined decision boundary and point that form the decision boundary between the classes called support vector treat as parameter. SVM is based on the machine learning algorithm invented by vapnik in 1960‘s. It is also based on the structure risk minimization principle to prevent over fitting.

There are 2 key implementations of SVM technique that are mathematical programming and kernel function. It finds an Optimal separates hyper plane between data point of different classes in a high dimensional space. Let's assume two classes for classification. The classes being P and N for Y_n= 1,-1, and by which we can extend to K class classification by using K two class classifiers. Support vector classifier (SVC) searching hyper plane. But SVC is outlined so kernel functions are introduced in order to non line on decision surface.

6.1. Linear SVC

Data that is linearly separable. Let w is weight vector, his base, X_n is the nearest data point $w^T+b\geq 1$ for x_n ε P and $w^T x_n+b\geq 1$ for x_n ε N.

For optimization the problem minimizes the ½ $[w^T w]$ Subject to $y_n (w^T w+b) \geq 1$ for n=1 to N.

6.2. Non –linear SVC

A linear classifier not suitable for c class hypothesis. It can be used to learn nonlinear decision function space SVM can also be extended for learning non-linear decision function.

6.3. Non Separable Case

Noise is present in the training data, some data point may be misclassified.

7. Role of Support Vector Machine (SVM) in Web Classification Optimized by Firefly

This section talks about the method, which are used to improve the performance of the web page classification method. This improvement is done by optimizing the basic feature selection method by Support Vector Machine (SVM). This is also called ensemble approach to classify web pages.

(1) Apply Firefly based feature selection process.

(2) Apply efficient Support Vector Machine (SVM) to classify webpage.

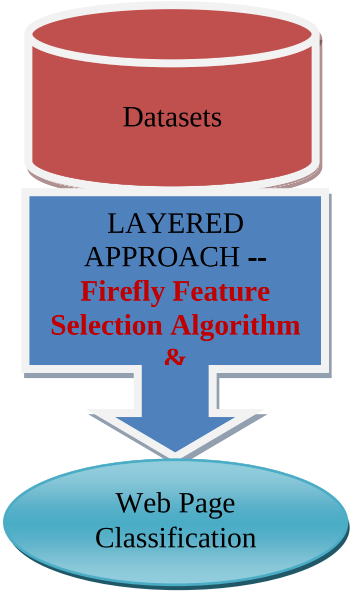

Figure 2. Architecture of Classifier based on Firefly and SVM

8. Conclusions and Future Work

Here we have seen different aspect like (i) Web classification, (ii) Optimization method (which is use by the web classification method to increase the efficiency or say decrease the complexity), (iii) Support Vector Machine. Our study found that what is web page classification and different ways to optimize it. It also gives the idea to extract various feature or information from various Web pages and use them according to their need or other way round, this extract information is used for web page classification.

This study motivates us to do further work in the area of optimized web page classification with the help of Support Vector Machine.

References

[1] C. Curtsinger, B. Livshits and B. Zorn, “ZOZZLE: Fast and recise In-Browser JavaScript Malware Detection”, SEC’11 Procedings of the 20th USENIX conference on Security. Berkeley, CA, USA: USENIX Association, **(2011)**, pp. 3-3.

[2] I. Fette, N. Sadeh and A. Tomasic, “Learning to detect phishing emails”, Proceedings of the International World Wide Web Conference (WWW), Banff, Alberta, Canada, **(2007)**.

[3] I. K. Murthy, “Data Mining- Statistics Applications: A Key to Managerial Decision Making”, SOCIO 2010, available at: http://www.indiastat.com/article/16/krishna/fulltext.pdf.

[4] M. H. Zack, “Developing a knowledge strategy: epilogue”, Available at: http:// web .cba.neu.edu/-mzack /articles. Y-Shapiro, P. Smyth, and R. Uthurusamy, 569–588. Menlo Park, Calif.: AAAI Press, **(2001)**.

[5] D. Hand, H. Mannila and P. Smyth, “Principles of Data Mining”, The MIT Press, 2001. Available at: ftp://gamma.sbin.org/pub/doc/books/Principles_of_Data_Mining.pdf.

[6] X. S. Yang, "Firefly algorithms for multimodal optimization", ochastic Algorithms: Foundations and Applications, SAGA 2009. Lecture Notes in Computer Sciences, vol. 5792, pp. 169–178.

[7] X.-S. Yang and S. Deb, "Cuckoo search via Levy flights", World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), IEEE Publication, USA, pp. 210–214.

[8] Novel Cuckoo Search Algorithm Beats Particle Swarm optimization http://www.scientificcomputing.com/news-DA-NovelCuckoo-Search-Algorithm-Beats-Particle-SwarmOptimization-060110.aspx, [last accessed on 25/7/2012].

[9] X.-S. Yang, "A New Metaheuristic Bat-Inspired Algorithm", Nature Inspired Cooperative Strategies for Optimization (NISCO 2010), Eds. J. R. Gonzalez et al., Studies in Computational Intelligence, Springer Berlin, 284, Springer, pp. 65-74.

[10] R. Tang, S. Fong, X.-S. Yang and S. Deb, "Wolf search algorithm with ephemeral memory", IEEE Seventh International Conference on Digital Information Management (ICDIM 2012), Macau, To appear, **(2012)** August.

[11] N. Jerne, "Towards a network theory of the immune system", Annals of Immunology (Paris), vol. 125, no. 1–2, **(1974)**, pp. 373–389.

[12] L. N. de Castro and J. Timmis, "Convergence and Hierarchy of aiNet: Basic Ideas and Preliminary Results", Proceedings of ICARIS (International Conference on Artificial Immune Systems), University of Kent at Canterbury, September 2002. University of Kent at Canterbury Printing Unit, pp. 31–240.

[13] L. N. de Castro and J. Timmis, "An Artificial Immune Network for Multimodal Optimisation", Congress on Evolutionary Computation, IEEE. Part of the 2002 IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, USA, **(2002)** May, pp. 699-704.

[14] O. M. Alonso, F. Nino and M. Velez, "A Robust Immune Based Approach to the Iterated Prisoner's Dilemma", G. Nicosia, V. Cutello, P. J. Bentley, and J. Timmis, editors, Proceeding of the Third Conference ICARIS, Edinburg, UK, **(2004)** September, pp. 290–301.

[15] V. Vapnik, "The nature of statistical learning theory", New York: Springer-Verlag, **(1995)**.

[16] P. Kaur, "Web Content Classification: A Survey", International Journal of Computer Trends and Technology (IJCTT), PP No ISSN: 2231-2803, vol. 10, no. 2, **(2014)** April.

[17] S. Singh Bisht and Prof. (Dr.) Sanjeev Ansal, "Optimization of Web Content Mining with an Improved Clustering Algorithm", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, vol. 3, no. 11, **(2013)** November, pp. 479-483.

[18] Sarac, Ozel, S. A., "Web Page Classification Using Firefly Optimization", IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), ISBN: 978-1-4799-0659-8 INSPEC Accession Number: 13710930, **(2013)** June 19-21, pp. 1-5.

[19] C. Zhang, J. Ning and D. Ouyang, "An artificial bee colony approach for clustering", Expert Systems with Applications, vol. 37, **(2010)**, pp. 4761-4767.

[20] D. Karaboga and C. Ozturk, "A Novel clustering approach: Artificial bee colony (ABC) algorithm", Applied Soft Computing, vol. 11, **(2011)**, pp. 652-657.

[21] X. Yan, Y. Zhu, W. Zou and L. Wang, "A new approach for data clustering using hybrid artificial bee colony algorithm", Neuro computing, vol. 97, **(2012)**, pp. 241-250.

[22] J. Kennedy, R. C. Eberhart and Y. Shi, Swarm Intelligence, Morgan Kaufmann, New York, **(2001)**.

[23] H. Shah-Hosseini, "Problem solving by intelligent water drops", Proceedings of IEEE Congress on Evolutionary Computation, Swissotel The Stamford, Singapore, **(2007)** September, pp. 3226-3231.

[24] http://en.wikipedia.org/wiki/Firefly_algorithm\.

[25] M. Kamber Han, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, CA, **(2006)**.

Authors

Shashank Dixit, he is research scholar in madhav institute of technology and science, Gwalior (M.P.) India under the supervision of Dr.R.K.Gupta. He has completed his bachelor degree in information technology from maharana pratap college of technology, Gwalior (M.P.) India and currently pursuing master of technology (M.Tech) degree in computer science .Research area of his interest includes data mining, data warehousing, text mining and web mining and classification techniques etc.

Dr. R. K. Gupta, he is working as head of the department of computer science and information technology in madhav institute of technology science, Gwalior (M.P.) India .he has received phd degree from ABV-IIITM gwalior (M.P.) India .he has post graduated (M.Tech) from IIT delhi India and he has held bachelor degree (B.E.) from madhav institute of technology and science Gwalior (M.P) India. He has many years of teaching experience and he has guided many Ph.D. students as well as M.Tech students. Numbers of research paper has been published by him in data mining .His area of interest is data mining, web mining etc.