International Journal of u- and e- Service, Science and Technology
Vol.8, No.5 (2015), pp.355-364
http://dx.doi.org/10.14257/ijunesst.2015.8.5.33

Layered Approach to Classify Web Pages using Firefly Feature Selection by Support Vector Machine (SVM)

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

[1]*Department of Computer Science and Engineering, MITS college, Gwalior, India*
[2]*Professor, Department of Computer Science and Engineering, MITS college, Gwalior, India*[1]*shashankdixit54@gmail.com,* [2]*iiitm_rkg@rediffmail.com*

Abstract

With the very fast development in today`s digital world, Information Retrieval on Internet is gaining importance, day by day. The web comprises of huge amount of data and search engine provides an efficient way to navigate the web and get the relevant information. The search engine has proven to be less efficient in providing relevant information from a query processed by a user. Fors olving this problem and getting accurate results there is need to categorize these web pages. Many optimizations have also done to speedup the classification process as it is required to be fast while maintaining the efficiency. To maintain the accuracy with the lesser time requirement, researchers have developed a SVM based Layered approach with the help of firefly feature selection method.

Keywords*: Web Page Classification, Firefly optimization, Support Vector Machine (SVM), Optimization Techniques*

1. Introduction

Through many years we have seen the growth in web pages that is increasing on WWW continuously. It is only accessible through search engines. WWW is the largest repositories of web page data and it is fully unordered and not well maintained. To organize large amount of data in well ordered and accurate way looks impractical by human efforts or doing it manually. There is need to order these data in efficient manner and in well structured format so that the web directory according to subject of web page can be maintained so it will make extraction of web page accurate and fast. For solving those problems there is a technique called web page classification by using this technique web pages can be categorized in specific categories [1]. It will not make only easier for user to browse web page but also easier to browse web page through search space.

Web page classification problem divided into two categories that are manual classification and automatic web page classification. Manual classification is a task that is performed by human being manually and it looks impractical because WWW contains millions of web pages and continuously increasing because it will take lots of human effort and time. While automatic web page classification is supervised machine learning problem where set of document is used to train the classifier once training is done it is used to classify web pages. It saves lot of manpower and material resources and time [2]. There are various machine learning techniques available in the literature such as Naïve bayes, K-nearest neighbor, Decision tree, neural network, support vector machine etc which have been used previously by many researchers to do this task.

For classification of web page accurately in respective class it is needed to identify the content of web pages. Once the content of web page is identified it can be decided easily that a web page belong to which category. Web page usually written in HTML form and those html tags are content of web pages they are also called feature of web page. Feature

ISSN: 2005-4246 IJUNESST

selection is important process for classification. Feature should be relevant, non redundant, noise free for accurate web page classification. There are some approach available for feature selection from web page are as follows filter approach, wrapper approach and swarm based optimization algorithm [3,4]. Filter approach is based on applying scoring method to evaluate effective feature from dataset of web pages for example document frequency, Chi Square and information gain etc. while wrapper approach wrap the feature around the classifier to classify data to anticipate the benefit of adding or removing certain feature from training data.

Swarm based optimization algorithm has inspired from nature and are very effective in selecting the best effective features from web page. There are various swarm based optimization algorithm are available some of them are Genetic Algorithm, Ant colony optimization, Particle swarm optimization, Intelligent water drop algorithm, Artificial Bee Colony and Firefly Algorithm etc. Firefly algorithm [7] is the recent and best search optimization algorithm. Now a day these are widely used with classification algorithm to optimize the performance of webpage classification. This paper proposes a novel scheme of feature discovery, through adding the measure of feature similarity of words in the discovery process using firefly optimization. It extends the focus of Feature discovery with detailed overview of Feature Mapping and classification concepts of support vector machine. The concept of support vector machine and firefly algorithm is discussed in next section.

The article is organized as follows: Section 2 present the working of support vector machine. Section 3 presents the concepts of firefly algorithm. Section 4 presents the related work. Section 5 presents the layered approach for classify web page using firefly feature selection by support vector machine. Section 6 shows the experimental setup and results. Section 7 presents conclusion and future work.

2.Support Vector Machine

The Support Vector Machine (SVM) is developed by Vapnik [5]. Generalization quality and effortlessness of training of SVM is remote away from the capacity of many conventional techniques. SVM model could be hard. Some real world troubles for example classification of the various text & image, recognition of handwriting, and analysis of various bioinformatics &bio sequence. SVM does very good with those data sets which have more and more attributes, yet here is less cases on that to train model. The support vector machine [SVM] is a training algorithm. It trains the classifier to predict the class of the new sample. The basis of the SVM is decision planes which clear decision boundary & spot that shape the decision boundary among various classes called support vector consider as one of the important parameter. It is also based on the structure risk minimization principle to prevent over fitting.

SVM algorithm originates in the processing of data classification problem, and is an implementation method of statistical learning theory. It is built on the basis of the limited number of samples, and it can get the best classification effect under existing training text. Algorithm also convert actual problem to high-dimensional feature space through nonlinear transformation. In the high-dimensional space it constructs a linear discriminate function to achieve the non-linear discriminate function in the original space. Its special nature ensures that the machine has a good generalization ability and cleverly figure out the problem of dimension. The complexity of the algorithm [10] is independent of the dimension of the sample. There are two types of support vector machine classifier

1.1.1 Linear SVM: Linear support vector machine classify data which is linearly separable in two classes and the SVM is selected with the maximum margin because of less generalization error and a slack variable is also used if data has noise.

$w^T x_n + b \geq 1$ *for* $x_n \varepsilon P$ *and* $w^T x_n + b \leq 1$ *for* $x_n \varepsilon N$

Where $w^T$is the weight vector and $x_n$is the nearest data point and b is bias (scalar value)

1.1.2 Non linear SVM: Data which are not separable by linear SVM due to high noise. Non linear SVM is used for separating data and data are mapped into high dimension space using kernel functions. There are various type of kernel functions which are used by SVM.

1. Linear kernel
2. Gaussian kernel
3. Gaussian (Radial-Basis Function (RBF)) kernel
4. Sigmoid Kernel

3. Firefly Algorithm

Xin-She Yang formulated firefly algorithm in 2008.Firefly Algorithm (FA) is a recent search and optimization techniqueIt is based on flashing behavior of fireflies [7]. Firefly communicate with each other via bioluminescent glowing which enables them to explore cost function space more effectively than in standard distributed random search. Each and every firefly gets attract towards other according to the flashing intensity or brightness of the firefly. The main rules of algorithm are as follows:

- All fireflies are of same gender (unisex).
- If there are more than two fireflies then a firefly with lesser of dull flashing light gets attracts towards more or higher flashing firefly.
- In the case when there is no firefly brighter then the given firefly then these all move randomly.

So moral of story, the objective function also play around flashing lights.Brightness of firefly can be calculated using this formula

$\beta = \beta_0 e^{\Upsilon r2}$

Where β0 is the brightness at distance r = 0 and ϒ is the light absorption coefficient.

4. Related Work

Firefly Algorithm brought into existence by Xin-She Yang in 2007-2008 at Cambridge University [7], this method has coined its idea from the nature and behavior of fireflies. There are lot many approaches proposed by a many researches in information retrieval. Firefly algorithm is a based wrapper technique which finds the best features for Web pages, to make fast and accurate classification. The algorithm constitutes a population-based iterative procedure with numerous agents (perceived as fireflies) concurrently solving a considered optimization problem. The working of this method used by different researchers to select feature from webpage is introduced below.

Due to the presence of the noisy data there is a need for classification of the web page for real world applications. A method which will properly ensure the classification is the support vector machine because it has the capability of generalization. However the accuracy measure of this method is not good because they are sensitive to noisy training data. Our suggested method provides a method which will increase the accuracy of

classification by combining the support vector machine concept with the K-nearest neighbor techniques. For any given set of training data this method first employs the KNN method so to remove the noisy data from this training data set. After that, the remaining training data are subjected to SVM for web page classification. Various simulation result have shown that these novel proposed method have strong resilience of noise and has the capability of decreasing the effect of noisy training data on support vector machine[10].

In 2013 EsraSaraç and Selma Ayşe Özel [7] have used firefly algorithm for classification of webpage. They have applied firefly algorithm on web kb dataset and conference dataset and selected number of feature (i.e html, xml tag) and then classified using j48 classifier of WEKA tool. They have also compared the performance in term of feature selection time and f-measure value with ant colony optimization and intelligent water drop algorithm. Firefly algorithm perform better it select 100 feature in 34.53 minutes in 250 iteration.

In 2011 J. Senthilnath, S.N.Omkar, V mani [11] have used Firefly Algorithm (FA) for clustering on benchmark problems and they have compared the performance with two nature inspired Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO). The classification error rate of firefly is lesser than the other two method. the FA achieved 11.36 while ABC, and PSO has achieved 13.13% and 15.99% respectively.

In research paper [12], researchers have combined the Rough Set Theory (RST) with nature inspired firefly algorithm and ant colony, bee colony, particle swarm optimization for feature selection. They have applied there new approach on four medical data sets namely cleveland heart dataset, lung cancer, Wisconsin, dermatology and compared the result with these techniques. The number of feature selected by firefly is i.e.3,5,4,7 with respective dataset which is less than the other techniques.

5. Layered Approach to Classify Web Pages using Firefly Feature Selection by Support Vector Machine (LACWPUFFS-SVM)

The main concept behind the proposed work is that it first finds all the features of the given dataset through firefly technique then apply SVM Classifier to classify web pages as shown in figure 2. The algorithm of proposed work is shown below in figure 1

Algorithm:

1. Start
2. Read Dataset of Web Pages
3. Find various possible features of All Web Pages using Firefly algorithm.
4. Apply SVM to divide the web Pages into categories.
5. Stop

Figure 1. Algorithm of Proposed Work

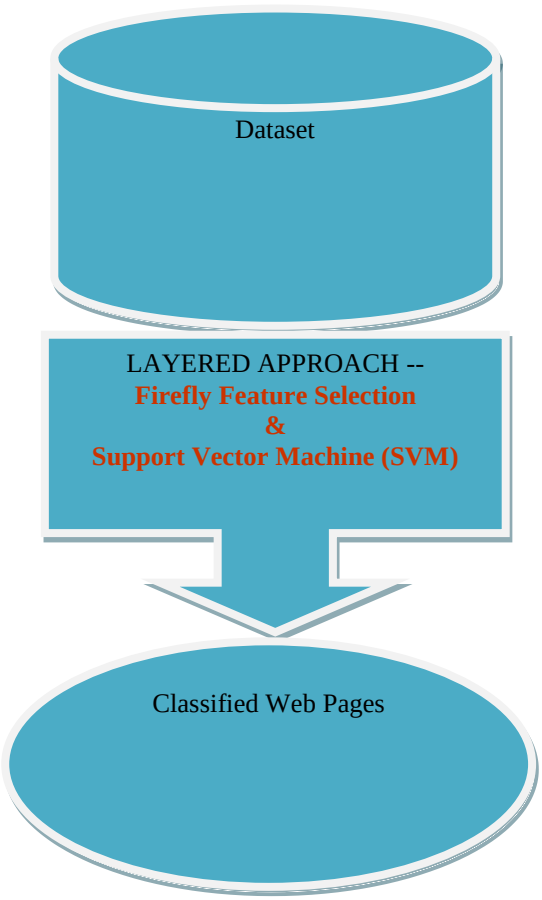


Figure 2. Architecture of Layered Approach to Classify Web Pages Using Firefly Feature Selection by Support Vector Machine (SVM)

6. Experimental Setup and Results

In this section, researchers have initiated with datasets used in the experiment.After that evaluated the performance of proposed work on datasets and execution result in term of classification accuracy and elapsed time is shown in following section.

Simulation Test

Dataset 1:

Researchers use a dataset which contains total 334 Web pages. These 334 Web pages are provided by Cornell University, Ithaca, NY: [8] for the research purpose. To perform test on datasets, Web pages are analyzed prior and divided into two groups i.e. course and Non-course. We have selected 224 webpage as course web page and 100 webpage as non course web pages. This method is works as follows: In first step by the method retrieves30 keywords from all pages of dataset. In the second step, researchers estimate category of the all Web pages watchfully. This estimation work is done by an automated software module. And then support vector machine is applied to classify webpage.

Dataset 2:

WebKBdataset were used in the experiments [7] too. It is well known data structure because it has been used previously many times for experiment purposes of web mining. It is freely available on web and we have taken total 1065 pages after we have divided in

two categories which are course and non course for experiment purpose. Course has contains 928 page while non course contains 137 pages. This method works in two steps: in first step we have selected 30 keywords from data set using firefly optimization and in second step we classify web page using support vector machine (SVM). It's various structures is shown in Table 1 and 2.

Table 1.WebkB Dataset

Category:	Non-course	Course
Type:	HTML files	HTML files
Quantity:	137	928

Table 2. Details Structure of Webkb Dataset

Category	Features	Quantity
Non-Course	Cornell	21
Non-Course	Texas	3
Non-Course	Washington	10
Non-Course	Wisconsin	12
Non-Course	Misc.	91
Course	Cornell	42
Course	Texas	38
Course	Washington	77
Course	Wisconsin	85
Course	Misc.	686

To execute LACWPUFFS-SVM and existing WPC-FF [7] methods, researchers are used Pentium Dual Core CPU G645 2.9Ghz with 4GB ram and window 7 with 64-Bit Operating System personal computer.

When researchers have executed the application, they found that classification results remains same but the execution time required to execute the application for the Proposed (LACWPUFFS-SVM) and web page classification with firefly optimization (WPC-FF)[7] becauseit changed in each execution. This is the reason to execute the application 8 repetitive times and to find the results. Later on, they are compared in Table 1, 2, 3 and 4 with their respective graphs 1, 2, 3 and 4.

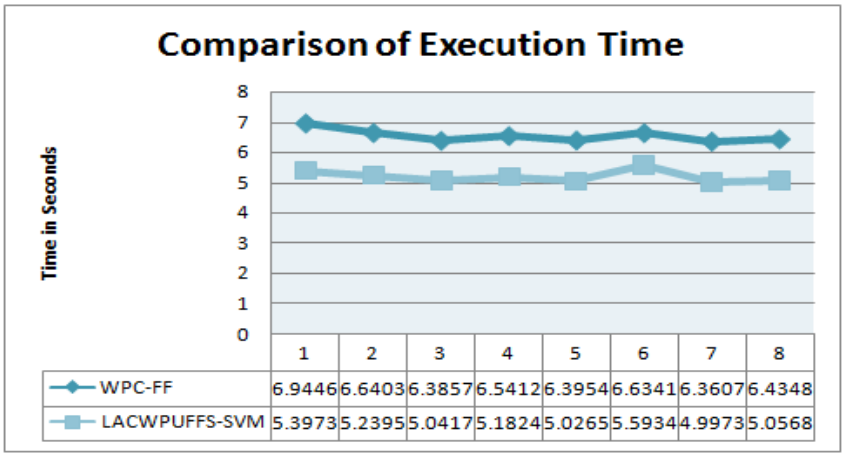
Table 1. Execution Time Comparison between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset-1

S. No.	WPC-FF	LACWPUFFS-SVM
1	6.94468	5.39736
2	6.64032	5.23957
3	6.38576	5.04178

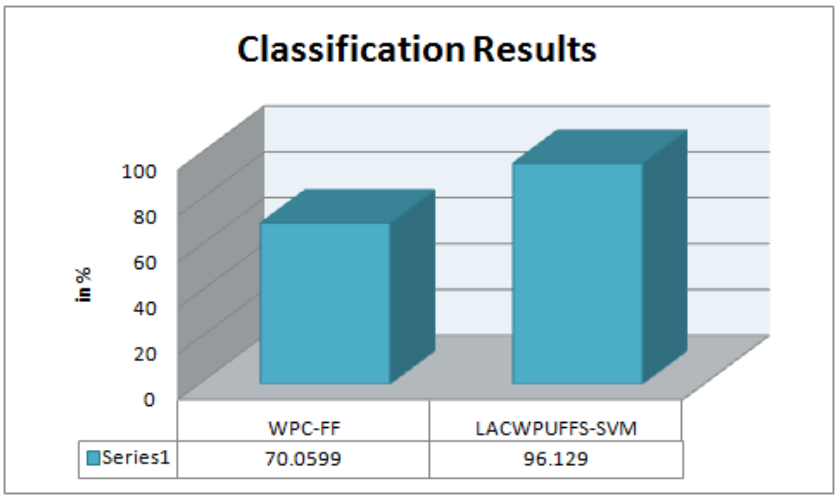
4	6.54129	5.18242
5	6.39542	5.02654
6	6.63416	5.59345
7	6.36077	4.99734
8	6.43483	5.05685

Table 2. Classification Results Comparison between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset-1

S. No.	WPC-FF	LACWPUFFS-SVM
1	70.0599	96.129



Graph 1: Comparison of Time of Execution between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on dataset-1



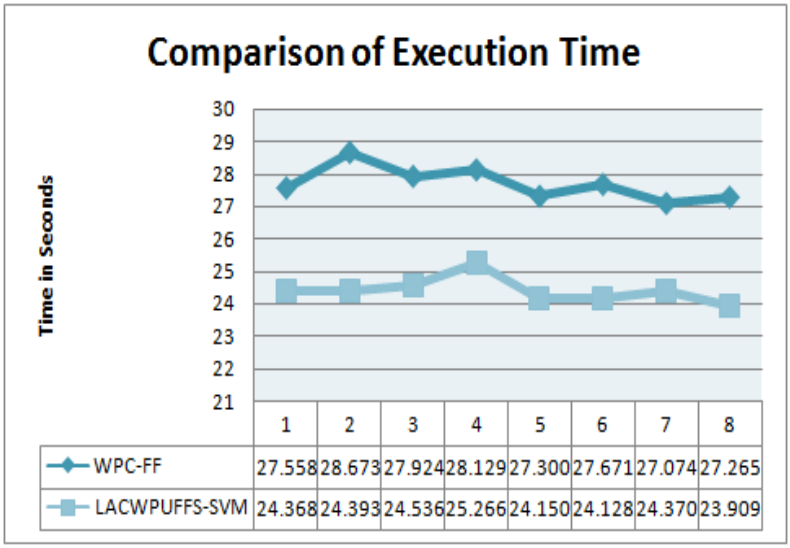
Graph 2: Comparison of Classification Result between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset 1

Table 3: Execution Time Comparison between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset-2

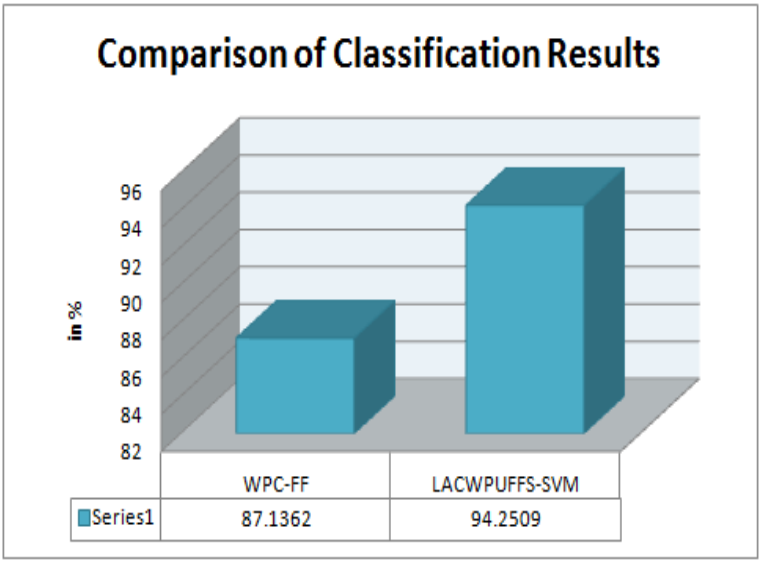
S. No.	WPC-FF	LACWPUFFS-SVM
1	27.5581	24.3689
2	28.6733	24.3939
3	27.9242	24.5367
4	28.129	25.2667
5	27.3001	24.1503
6	27.671	24.1284
7	27.0742	24.3705
8	27.265	23.9093

Table 4: Classification Results Comparison between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset-2

S. No.	WPC-FF	LACWPUFFS-SVM
1	87.1362	94.2509



Graph 3: Comparison of Time of Execution between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7]on Dataset-2



Graph 4: Comparison of Classification Result between Proposed (LACWPUFFS-SVM) and Existing Work (WPC-FF)[7] on Dataset 2

7. Conclusion and Future Work

In this research we have developed a layered approach to classify webpageusing firefly feature selection by support vector machine. We have also compared proposed (LACWPUFFS-SVM) method with firefly feature selection (WPC-FF) method in terms of classification accuracy and time. Here we have selected best feature i.e. html tag from web pages using firefly algorithm so that dimensionality of features get reduced and after that on the basis of selected feature we have applied support vector machine for classification. Eventually, our developed method achieved 96 percent classification accuracy at the same time required for execution is lesser then other method.

In future different classification techniques and optimization techniques can be used for classification of web pages and experiment can be performed on other data sets also.

References

[1] H. Yu, J. Han and K.C. C. Chang, "PEBL: Positive Example Based Learning for Web Page Classification using SVM", Proc. ACM SIGKDD International conf. Knowledge discovery in Databases (KDD), ACM Press, **(2012)**; New York.
[2] D. Zhang and W. S. Lee, "Web taxonomy integration using support vector machine", Proceedings of the 13th international conference on World Wide Web, ACM Press, **(2004)**; New York, USA.
[3] M. Aghda, N. G. Aghaee and M. Basiri, "Text feature selection using ant colony optimization", Expert Systems with Applications, vol. 8, no. 22, **(2008)**.
[4] C. Chen, H. Lee and Y. Chang, "Two novel feature selection approaches for web page classification", Expert System with Applications, vol. 36, **(2009)**, pp. 260-272.
[5] V. Vapnik, "The nature of statistical learning theory", **(1995)**; New York, Springer-Verlag.
[6] X. S. Yang, "Firefly algorithms for multimodal optimization", Stochastic Algorithms: Foundations and Applications, SAGA, Lecture Notes in Computer Sciences, vol. 5792, **(2009)**, pp. 169–178.
[7] E. Saraç and S. A. Özel, "Web Page Classification Using Firefly Optimization", Innovations in Intelligent Systems and Applications (INISTA), IEEE International Symposium, **(2013)**.
[8] Computer Science Department, Cornell University, **(1996)**; Ithaca, NY.
[9] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K Nigam and S. Slattery, "Learning to extract symbolic knowledge from the World Wide Web", The 15th National Conference on Artificial Intelligence, AAAI Press, **(1998)**.
[10] J. Q. Zou, G. L. Chen and W. Z. Guo, "Chinese Web page classification using noise-tolerant support vector machines", Natural Language Processing and Knowledge Engineering, IEEE NLP-KE, **(2005)**, pp. 785 - 790.

[11] J. Senthilnath, S. N. Omkar and V. Mani, "Clustering using firefly algorithm: Performance study", Swarm and Evolutionary Computation, vol. 1, no. 3, **(2011)**, pp. 164–171.
[12] H. Banati and M. Bajaj, "Fire Fly Based Feature Selection Approach", IJCSI International Journal of Computer Science Issues, vol. 8, no. 2, **(2011)**.

Authors

Shashank Dixit, he is pursuing M. Tech in computer science from Madhav Institute of Technology and Science (M.P.) Indiaunder the supervision of Dr. R. K. Gupta.He has completed bachelor degree in information technology from MaharanaPratap College of Technology Gwalior (M.P.) India. Research areas of his interest are web mining, Data warehousing, text mining and classification techniques, optimization techniques etc.

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 been 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 papers have been published by him in data mining. His areas of interest are data mining, web mining etc.

