International Journal of Advance Foundation and Research in Computer (IJAFRC)
Volume 1, Issue 9, September 2014. ISSN 2348 – 4853.

A Review on Optimization in Web Page Classification.

Nikita Sahu, Dr. R. K. Kapoor
Department of Computer Engineering & Application, NITTTR, Bhopal M.P., INDIA
Associate Professor, Department of Computer Engineering & Application, NITTTR, Bhopal M.P., INDIA
nikita.sahu@ymail.com, rkkapoor@nitttrbpl.ac.in

A B S T R A C T

The rapid development of the internet and web publishing techniques create numerous information sources published as HTML pages on World Wide Web. WWW is now a popular medium by which people all around the world can spread and gather the information of all kinds. But web pages of various sites that are generated dynamically contain undesired information also. This information is called noisy or irrelevant content. The need for innovative and effective technologies to help find and use the useful information and knowledge from a large variety of data sources is continually increasing. Web information has become increasingly diverse. In order to utilize the Web information better, people pursue the latest technology, which can effectively organize and use online information. Classification is one of the vital and important data mining techniques that grouped various items in a collection to predefined classes or groups. The main goal of classification is to exactly predict the target class for each case in the data. Web Page Classification is technique of data mining to discover classification of web pages. The information providers on the web will be interested in techniques that could improve the effectiveness of the web search engine. In this paper, the relationships among the techniques used in data mining are studied. A study of web usage is also done on optimization of this web classification.

Index Terms - Web Mining, Classification, Data Mining Techniques, Optimization, Web Page, Danger Theory, Artificial Immune System.

I. INTRODUCTION

Over the past decade, web users have witnessed an exponential growth in the number of web pages accessible through popular search engines. Organizing the large volume of web information in a well-ordered and accurate way is critical for using it as an information resource. One way of accomplish this in a meaningful way requires web page classification. Web page classification addresses the problem of assigning predefined categories to the web pages by means of supervised learning. This inductive learning process automatically builds a model over a set of previously classified web pages. The learned model is then used to classify new web pages. Numerous classifiers proposed and used for machine learning can be applied for web page classification. These include Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN), and Naïve Bayes (NB) classifiers.

The most common form of malicious web pages is that containing virus and Trojan [1], which make use of the security vulnerability of the browser and operating system to attack users' computer system. Phishing [2] is another form of malicious web pages. Its main purpose is to cheat personal or financial information of Internet users. In addition, malicious web advertisement is becoming more and more popular.

The attack characteristic of malicious web pages is that it spreads on the Internet rapidly and widely with web pages as carrier. Generally, a malicious web page takes a passive mode to attack the user's computer system when a user browses the web page. A few malicious web codes may attack the web page with the security vulnerabilities by search engines. At present, most of the malicious web codes take active attack. Once the web server is attacked and infected with malicious code, it will serve as a malicious web page server. When the users browse the web pages in the malicious web server, their computers are likely to be infected by malicious programs.

WWW users want to find desirable many and only web pages from the vast numbers of web pages on WWW through web search engines. General web search engines generate search results based on correspondence between query phrases and the phrases in web pages. Such search engines do not consider an individual user's phrase meaning. As an example,

Suppose that WWW users A and B search for web pages by the query phrase "Web application". On the other hand, client B does with the meaning "Web mail" and "Web-based office software". The search result has included the web pages which is not relevant to the query phrase in each user's mind, which is undesirable for users.

A large number of statistical learning methods have been applied to the text classification problem in recent years. Some of them are regression models, nearest neighbor classifiers, Bayesian probabilistic classifiers and decision trees, inductive rule learning algorithms, neural networks and on-line learning approaches.

The structure of rest part of the article is structured as bellows: Section 2 briefly introduces various types of data mining techniques. Section 3 talks about web data mining and the web page classification process in brief. Section 4 concentrates and provides various techniques of optimization in data mining, especially in web page classification. Section 5 talks about Danger Theory concept of Artificial Immune System. Section 6 talks about the output of this study in terms of conclusion.

II. DATA MINING TECHNIQUES

Data mining is a process to extract interesting, implicit, previously unknown and potentially useful knowledge or patterns from data in large databases [3]. It is one of those latest technology with potential to support or help organizations on the very vital processed data in their respective warehouse of data. Each and every Data mining technique is output of exhaustive research and development work. This growth always get start when data of organization's business stored on computers, already, and still progressing with lots of improvements in accessing of data, and added all updates recently, generated technologies which permit users to work and navigate through their respective data in real time scenario. With recent progress in automated data gathering and the availability storage, a lot of businesses have routinely started collecting massive amounts of data on various facets of the organization. The eventual goal of this data gathering is to be able to use this information to gain a competitive edge by discovering previously unknown patterns in the data that could guide the process of decision making. Tools of Data mining could be able to answer all questions related to business which originally were very much time consuming to get resolve. Various Queries such as "Which clients are most likely to respond to my subsequently promotional mailing, and why?"

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 [4]. 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) [5,6]. 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 [4].

III. WEB DATA MINING AND WEB CLASSIFICATION

World Wide Web (WWW) has been proving to be tremendous amount of data and also data on WWW is growing exponentially in terms of both their size and its usage with respect to time. In contrast to the standard data mining methods web data mining methods need to deal with heterogeneous, semi structured or unstructured data [7]. In Web Data Mining various core or applied data mining techniques are applied to obtain some interesting knowledge out of data available on WWW. Also the resources (web pages) on WWW undergo frequent updation in terms of their content, structure, with respect to time. Web data mining can be categorized based on the interest and/or final objective of what kind of knowledge to mine from web data [8]. 1) Web Content Mining: refers to discovery of useful information or knowledge from web page contents i.e. text or it could be multimedia data like image, audio, video etc. 2) Web Structure Mining aims at analyzing, discovering and modelling link structure of web pages and/or web site to generate structural summary on which various techniques are applied and outcomes of these techniques can be utilized to recreate, redesign the web site which ultimately improves structural quality of web site [9]. 3) Web Usage Mining deals with understanding of user behavior, while interacting with web site, by using various log files to extract knowledge from them. This extracted knowledge can be applied for efficient reorganization of web site, better personalization and recommendation, improvement in links and navigation, attracting more advertisement. As a result more users attract towards web site hence will be able to generate more revenue out of it [8, 9, 10].

The working of WUM has three steps – first of all, pre-processing of the data which is to be used for find various classes, 2nd of all, pattern discovery and last but not least, analysis of the patterns. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources are not only discovered quality patterns but also improve the WUM algorithms. Hence, data pre-processing is very critical and important activity for the complete web usage mining processes and it plays a vital role in deciding the quality of patterns. In data pre-processing, the collection of various types of data differs with each other not only in type of data available but also the data source sites, the data source size and the way it is being implemented. While working with any Web mining, one should consider all these.

IV. 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.

A. **Firefly algorithm:** It is a meta heuristic algorithm. This algorithm is motivated by behaviour of fireflies [11]. To attract other fireflies is the main aim of firefly's flash. Xin-She Yang considered the firefly optimization by following factors: 1. Every fireflies are of same gender or say single

gender 2. Level of brightness of fireflies make them attractive according to brightness, It means among various fireflies, the lesser bright firefly will attract (and thus move) to the higher bright firefly 3. In rest of the cases when there is no firefly brighter than a given firefly, it will move randomly. So objective function must have brightness component. Current study have shown that Firefly Algorithm is mainly appropriate for nonlinear problems which follows multimodal.

B. **Cuckoo Search:** It is another algorithm of optimization which is developed by Xin-She Yang and Suash Deb in 2009 [12]. It was dully motivated by the brood parasitism of cuckoo species. This is done by laying their respective eggs in another host birds' nest. Out of all host birds, some could hold straight divergence with the pushy cuckoos. Some out of all cuckoo species have become part in the way that female of species of parasitic cuckoos are very particular in the imitation specially in various colors and pattern of thier eggs of host species which are chosen. Out of so available optimization method this Cuckoo Search proposed such behavior works of breeding at almost of all spaces. Cuckoo Search finds the various expressions: Where each egg shows a solution in particular nest and so each egg is nothing but a solution provided by cuckoo search. The primary and most important aim of egg replace provides most optimized solutions. We consider one egg per nest for our convenient. From the study of various optimization technique in various area, it is very clear that cuckoo search perform better than any colony, particle swarm optimization, etc. It's success make is applicable for more tedious and complex cases. This case could be consider that every nest has more than one cuckoo egg. "Novel 'Cuckoo Search Algorithm' perform outstanding than Particle Swarm Optimization". This was recently after comparison of various techniques in different area application [13].

C. **Bat Algorithms:** It is again very effective optimization technique which is based on a meta heuristic search method which is innovated by Xin She Yang in 2010 [14]. Echolocation behaviour of microbat is responsible for this algorithm. This is with changeable pulse emission with loudness. We can be summarized idolization of echolocation as below: Where we can consider that virtual bat flies randomly (velocity vel_i), which is at position x_i (which we can consider as solution). This is always consider along varying frequency A_i at any ith step of process.

D. **Wolf Search:** Wolf Search uses algorithm based on meta heuristic [15]. Main motivation behind it the wolves' hunting behavior which move as a pack. Here we know that each and every individual searching process of agent hunts. It is for a prey separately, mutely, which means without any further communication, and It is very obvious that they get merge by simply stirring their current locations to their respective peers' locations if and only if when the new terrains are performed much better than the one which is old. Here it is very important to state that mostly all wolves have their own visual range and only be in motion in levy journey. This all happens in the process of searching food [15].

V. DANGER THEORY IN ARTIFICIAL IMMUNE SYSTEM

This section evaluates related works in artificial immune network field in part-A and danger theory concept in Section B.

a. Artificial Immune Network

Artificial immune networks (AIN) are based on the immune network theory proposed by Jerne [16]. In 2001 de Castro and Von Zuben proposed this model for data analysis tasks. Their model it generates a

network of antibodies linked according to the affinity (Euclidean distance). A subset of the antibodies with the maximum affinity, with respect to a given antigen, is selected and clone proportionally to the affinity.

De Castro and Timmis [17] in 2002proposed a stopping criterion for aiNet algorithm based on Minimal Spanning Trees that is named Hierarchy of aiNets. It is possible to separate automatically the clusters, and sub-clusters, found in training data sets.

De Castro and Timmis [18] in 2002 proposed opt-aiNet. In This model the network cells interact accordingly with its affinity and by a suppression process that consists of removing those cells which affinities are less than a fix threshold. Otherwise, the cells go on cloning and mutation processes.

Alonso et al. [19] make a modification of aiNet to model an agent that plays the Iterated Prisoner's Dilemma (IPD) that try to find a strategy (most stimulated B-cell) in the immune memory. The main modification made to aiNet is in the memory mechanism: if a B cells added to memory it will never be removed. In this paper we concentrate on immune network algorithms as a main branch of artificial immune systems for anomaly detection to simulate adaptive immune system of our proposed method.

b. Danger theory

In general case, there are two generations of artificial immune system. One of this, only deals to simulate adaptive immune system, but another one which is called danger theory simulate both adaptive and innate immune system simultaneously.

One usage of danger theory is named dendritic cell algorithm that proposed by Greensmith, Aickelin [20]. DCA attempts to simulate the power of DCs which are able to activate or suppress immune responses by the correlation of signals representing their environment, combined with the locality markers in the form of antigens [21]. Another one is named toll-like receptors algorithm (TLR) was proposed by Twycross, Aickelin [22].

The DCA relies on the signal processing aspect by using multiple input and output signals, while the TLR emphasizes the interaction between DCs and T cells, only uses danger signal [18].

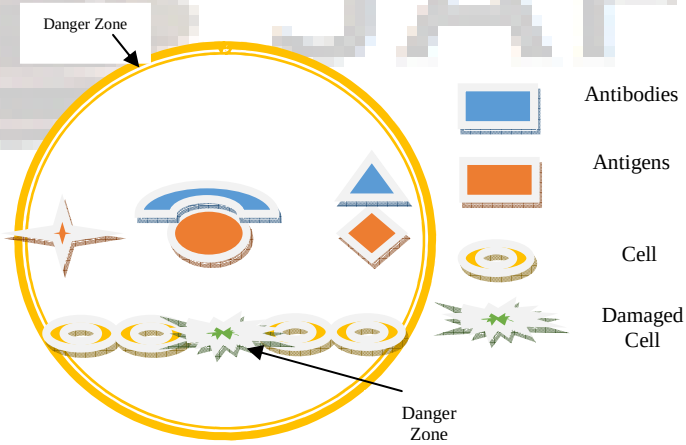

Figure 1. Danger Theory Model [23]

Figure 1 depicts how we might picture an immune response according to the Danger Theory. Essentially, the danger signal establishes a danger zone around itself.

In summary, both DCA and TLRA employ the model of DCs, which is an important element in the innate immune system [24]. However, DCA disregard the adaptive immune system but TLR employ the model of adaptive immune system by using self/non-self mechanism. In This work, we concentrate on the combination of immune network and k nearest neighbor classifier to present a novel method in danger theory field Proposed Method.

VI. CONCLUSIONS AND FUTURE WORK

This paper has discussed about three areas i) Web classification, ii) Optimization method (which is used by the web classification method to increase the efficiency or say decrease the complexity), iii) In the last, Danger Theory (Which could directly reject those web pages which behaves abnormally) has been discussed . This study will certainly be helpful for evaders to have a deep insight in to web page classification and different ways to optimize it. This study motivates us to do further work in the area of optimized web page classification with the help of danger theory.

VII. REFERENCES

[1] Charlie Curtsinger, Benjamin Livshits, Benjamin Zorn etal. ZOZZLE: Fast and recise In-Browser JavaScript Malware Detection. In: SEC'11 Procedings of the 20th USENIX conference on Security. Berkeley, CA, USA: USENIX Association, 2011 3-3.

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

[3] J. Han and M. Kamber, "Data mining: Concepts and techniques," China Machine Press, vol. 8, pp. 3-6 2001.

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

[5] Zack, M.H. " 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.

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

[7] Qingyu Zhang, Richard Segall "Web mining: a survey of current research, techniques and software", International Journal of Information Technology & Decision Making Vol. 7, No. 4, 2008.

[8] Kosala and Blockeel, "Web Mining Research: A Survey", SIGKDD Exploration, Newsletter of SIG on Knowledge Discovery and Data Mining, ACM, Vol.2, 2000.

[9] B. Singh, H. K. Singh, "Web Data Mining Research: A Survey", IEEE, 2010.

[10] J. Srivastav, R. Cooley, M. Deshpande, P. Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Vol.7, No.2, Jan-2000.

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

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

[13] "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].

[14] Yang X.-S., "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.

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

[16] N. Jerne, towards a network theory of the immune system, Annals of Immunology (Paris) 125 (1–2) (1974) 373–389.

[17] L. N. de Castro and J. Timmis. Convergence and Hierarchy of aiNet: Basic Ideas and Preliminary Results.InProceedings of ICARIS (International Conference on Artificial Immune Systems), pages-231– 240. University of Kent at Canterbury, September 2002. University of Kent at Canterbury Printing Unit.

[18] L. N. de Castro and J. Timmis. An Artificial Immune Network for Multimodal Optimisation. In Congress on Evolutionary Computation, IEEE. Part of the 2002 IEEE World Congress on Computational Intelligence, pages699 – 704, Honolulu, Hawaii, USA, May 2002.

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

[20] J. Greensmith, U. Aickelin, Dendritic cells for real-time anomaly detection, in: Proceedings of the Workshop on Artificial Immune Systems and Immune System Modeling (AISB'06), Bristol, UK, (2006), pp.

[21] J. Greensmith, U. Aickelin, G. Tedesco, Information fusion for anomaly detection with the dendritic cell algorithm, Information Fusion 11 (1) (2010).

[22] J. Twycross, U. Aickelin, Detecting anomalous process behaviour using second generation artificial immune systems. Retrieved 26 January 2008, from http:// www.cpib.ac.uk/jpt/papers/raid-2007.pdf, 2007.

[23] Uwe Aickelin, Steve Cayzer, "The Danger Theory and Its Application to Artificial Immune Systems," Proceedings of the 1st Internat Conference on ARtificial Immune Systems (ICARIS-2002), pp 141-148, Canterbury, UK, 2002.

[24] Shelly Xiaonan Wu*, Wolfgang Banzhaf, The use of computational intelligence in intrusion detection systems, Applied Soft Computing 10 (2010).

