International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 4, July-August 2014 **ISSN 2278-6856**

Review: Optimized Webpage Classification

Rupali A. Mulay [1], Abhishek Singh Chauhan [2]

[1]Department of Computer Science & Engineering,
NIIST RGPV University, INDIA

[2]Astt. Professor, Department of Computer Science & Engineering,
NIIST RGPV University, INDIA

Abstract

Now, with the explosive growth of the data stored in various forms, 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. Somehow we could put these all into this way as well like, 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. However, semi-structured document differ from tradition text which is neat and clean; 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 information providers on the web will be interested in techniques that could improve the effectiveness of the web search engine.

Keywords: Web Mining, Classification, Data Mining Techniques, Optimization.

1. Introduction

In recent years, with the fast development of the Internet, a large number of companies provide rich network resource and application services by building up their own websites. Even if the computers which have antivirus software installed can prevent most malicious attacks, various types of malicious web pages, such as phishing, malware and spamming, can cause a huge information and property damage to innocent users. 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, User B does with the meaning "Web mail" and "Web-based office software". The search result includes the web pages which is not relevant to the query phrase in each user's mind, which is undesirable for users. In this paper, the relationships among the techniques and data mining, Web usage are studied and optimization of this web 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 conclusion of the review.

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 has three distinct phases involved – content, structure and usage mining of web data. Mining the content involves extracting the relevant information, structure mining studies the structure and prototype and usage mining is

the analysis of the discovered patterns. Web Usage Mining (WUM) is all about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc. 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.

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 behaviour 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 technique 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 Algorithms: 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 behaviour of microbats 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 Wolf Search: It is one of the most recent meta heuristic algorithms [10]. It is inspired by the hunting behavior of wolves that move as a pack; each individual searching agent hunts for a prey individually, silently (without any communication) and they merge by moving their current positions to their peers' positions if the new terrains are better than the old ones. Wolves have certain visual range and move in levy flight in the food searching mode. A random hunter is implemented from which a wolf will jump out of its current visual range to a random position upon encounter. This random escape enables the wolves to stay out of a local subspace. Wolf search algorithm was shown to be more superior to the existing bio-inspired algorithms in [10].

5.Danger Theory In Artificial Immune System

We evaluate related works in artificial immune network field in Section A and danger theory field in Section B.

5.1 Artificial Immune Network

Artificial immune networks(AIN) are based on the immune network theory proposed by Jerne [11]. 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 highest affinity, with respect to a given antigen, is selected and cloned proportionally to the affinity. De Castro and Timmis [12] 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 [13] 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. [14] 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.

5.2 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[15]. 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 [16]. Another one is named toll-like receptors algorithm (TLR) were proposed by Twycross ,Aickelin [17]. 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 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 [18]. 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.

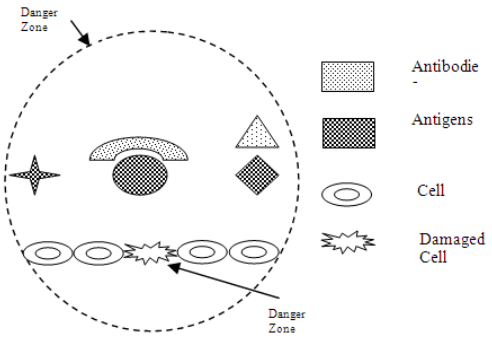

Figure 1 Danger Theory Model [21]

6. Role Of Danger Theory In Web Classification Optimized By Firefly

All methods of abnormal behaviour detection involve the gathering and analysis of information from various areas within a computer or network to identify possible threats posed by hackers and crackers inside or outside the organization. In this area, IDSs fall into two categories according to the detection approaches they employ, namely i) anomaly detection and ii) Misuse detection. Misuse detection identifies intrusions by matching observed data with pre-defined descriptions of intrusive behaviour. Therefore, well-known intrusions can be detected efficiently with a very low false alarm rate. But this approach will fail easily when facing unknown intrusions. Anomaly detection is orthogonal to misuse detection. It hypothesizes that abnormal behaviour is rare and different from normal behaviour. Hence, it builds models for normal behaviour and detects anomaly in observed data by noticing deviations from these models. Clearly, anomaly detection has the capability of detecting new types of intrusions, and only requires normal data when building profiles. Main point is to differentiate between normal and abnormal pattern [19]. Natural or say our immune system work for protecting our self from various outsider unwanted bodies like viruses and bacteria[20]. System is made up of various things like abilities, including pattern recognition, inherent distributed parallel processing etc. There are variety of techniques available which could be useful to simulate the immune system that are aimed at resolve the concern at very large extend. Common techniques in this field that explain behaviour of the immune system are Clonal Selection, Negative Selection, Immune Network and danger theory algorithms.

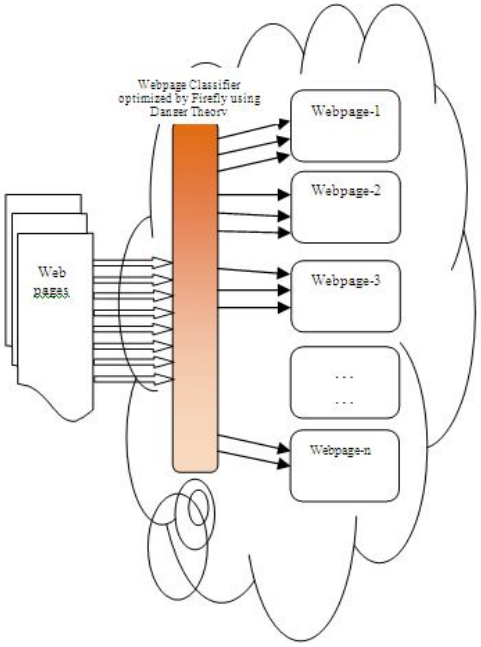

Figure 2 Architecture of classifier based on firefly and Danger Theory

Danger theory methods simulate both innate and adaptive immune system simultaneously so in this paper,

we propose a new method for anomaly detection which is inspired from danger theory and simulate adaptive and innate immune systems using immune network and k nearest neighbor classifier respectively. The same concept can be useful for finding the abnormal web pages in classification when we do web classification.

7. 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) Danger Theory (Which could directly reject those web pages which behaves abnormally).Our this study found that what is web page classification and different ways to optimize it. This study motivate us to do further work in the area of optimized web page classification with the help of danger theory. This concept is show by following picture.

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.] 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.] 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.

[4.] 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.

[5.] 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.

[6.] 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.

[7.] 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.

[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.] 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.

[10.] 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.

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

[12.] 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.

[13.] 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.

[14.] 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.

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

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

[17.] 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.

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

[19.] S.X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review", Applied Soft Computing, vol. 10, pp. 1–35, 2010.

[20.] L. N. de Castro and J Timmis, Artificial Immune Systems: A New Computational Intelligence Approach, 2002.

[21.] 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

AUTHOR

Rupali A. Mulayappered for M.Tech in Computer Science & Engg from Ragiv Gandhi Proudyogiki Vishwavidyalay University & received the B.E. degrees in Inforamtion Teechnology from Shri Sant Gadge Maharaj University in 2011.Presenting 2nd International Conference on ICIE 2013.

Abhishek Singh Chauhan (Asst. Prof.) appeared PhD. [Computer Science & Engineering] from Bhagwant University, Ajmer (Rajasthan) & received M.Tech from Rajeev Gandhi Prodyogiki Vishwavidhyalaya, Bhopal (2012), MCA from IGNOU (2005), M.Com from Pt. Ravishankar Shukla University, Raipur(2000), PGDCA from Makhanlal Chaturvedi University, Bhopal (2001), GNIIT fromNIIT (Graduate from National Institute of Information Technology) (1999). A Comprehensive Survey on Frequent Pattern Mining from Web Logs. (IJAEA Jan 2011).A Novel Approach for web usage mining using Growing Neural Gas. (National Conference on Recent Trends in Mathematics & Computing (RTMC) April 2011).Detecting and Searching System for Event on Internet Blog Data Using Cluster Mining Algorithm (Springer) http://www.springerlink.com/content/j13834wp348211j0/ (Jan 2012).Rule Based Lexical Dictionary based polarity analysis for reviews, International Journal of Emerging Trends in Electronics & Computer Science (Jun 2013).Secure Content Sniffing For Web Browsers : A Survey, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.Review on Classification of Web Log Data using CART Algorithm, International Journal of Computer Applications (0975 – 8887) Volume 80 – No 17, October 2013.Mining Association Rules from Infrequent Itemsets: A Survey, International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol. 2, Issue 10, October 2013, (5801-5808).