Natural Computation for Business Intelligence from Web Usage Mining

Ajith Abraham

IITA Professorship Program, School of Computer Science and Engineering, Chung-Ang University, Korea, ajith.abraham@ieee.org

Abstract

Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on. This paper introduces the various ingredients of natural computation and further presents the important concepts of Web usage mining and its various practical applications.

1. Introduction

The WWW continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW. Based on several research studies we can broadly classify web mining into three domains: content, structure and usage mining. Web content mining is the process of extracting knowledge from the content of the actual web documents (text content, multimedia etc.). Web structure mining is targeting useful knowledge from the web structure, hyperlink references and so on. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web [7]. Web usage mining has become very critical for effective web site management, creating adaptive web sites, business and support services, personalization and network traffic flow analysis [16][18][19].

Natural computation involves ideas/themes inspired from the nature including biological, ecological and physical systems. It is an emerging interdisciplinary area in which a range of techniques and methods are studied for dealing with large, complex, and dynamic problems.

Natural computation and its key paradigms are introduced in Section 2. In Section 3, web usage mining related research is summarized. Further in Section 4 a framework for web usage mining is presented and some conclusions are also provided towards the end.

2. Natural Computation and its Ingredients

Natural Computation is a general term referring to computing inspired by nature. The idea is to mimic the complex phenomena occurring in the nature as computational processes in order to enhance the way computation is performed mainly from a problem solving point of view. Characteristic for man-designed computing inspired by nature is the metaphorical use of concepts, principles and mechanisms underlying natural systems. The general area of computational intelligence is currently undergoing an important transformation by trying to incorporate computational ideas borrowed from the nature all around us:

- Artificial immune system
- Swarm intelligence
- Evolutionary computation
- Neural computation
- Bacterial foraging
- DNA computation
- Computing with words
- Simulated annealing
- Cultural algorithms
- Membrane computing
- Quantum computation
- Artificial life
- Hybrid approaches

2.1 Artificial Neural Networks

Artificial neural networks have been developed as generalizations of mathematical models of biological nervous systems. In a simplified mathematical model of the neuron, the effects of the synapses are represented by *weights* that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function, which is usually the sigmoid, Gaussian function etc. The neuron impulse is then computed as the weighted sum of the input

signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. The learning situations in neural networks may be classified into three distinct sorts. These are supervised learning, unsupervised learning and reinforcement learning [28].

2.2. Evolutionary Algorithms (EA)

Evolutionary algorithms are adaptive methods, which may be used to solve search and optimization problems, based on the genetic processes of biological organisms [29]. Over many generations, natural populations evolve according to the principles of natural selection and 'survival of the fittest'. By mimicking this process, evolutionary algorithms are able to 'evolve' solutions to real world problems, if they have been suitably encoded. Usually grouped under the term evolutionary algorithms or evolutionary computation, we find the domains of genetic algorithms, evolution strategies, evolutionary genetic programming and learning programming, classifier systems. They all share a common conceptual base of simulating the evolution of *individual* structures via processes of *selection*, *mutation*, and *reproduction*.

2.3. Swarm Intelligence

Swarm intelligence is aimed at collective behaviour of intelligent agents in decentralized systems. Most of the basic ideas are derived from the real swarms in the nature, which includes ant colonies, bird flocking, honeybees, bacteria and microorganisms etc. Swarm models are population-based and the population is initialised with a population of potential solutions [30]. These individuals are then manipulated (optimised) over many several iterations using several heuristics inspired from the social behaviour of insects in an effort to find the optimal solution. Ant Colony Optimization (ACO) algorithms are inspired by the behavior of natural ant colonies, in the sense that they solve their problems by multi agent cooperation using indirect communication through modifications in the environment. Ants release a certain amount of pheromone (hormone) while walking, and each ant prefers (probabilistically) to follow a direction, which is rich of pheromone. This simple behavior explains why ants are able to adjust to changes in the environment, such as optimizing shortest path to a food source or a nest. In ACO, ants use information collected during past simulations to direct their search and this information is available and modified through the environment.

2.4. Artificial Immune System (AIS)

The artificial immune systems like other biologically inspired techniques, tries to extract ideas from a natural system, in particular the vertebrate immune system, in order to develop computational tools for solving engineering problems. The basic idea of AIS is to exploit the immune system's characteristics of learning and memory to solve a problem. AIS can be broadly categorized into three subgroups: those using the clonal selection theory, those using negative selection and those using the immune network theory as their main inspiration [32].

2.5. Bacterial Foraging

Selection behavior of bacteria tends to eliminate animals with poor foraging strategies and favor the propagation of genes of those animals that have successful foraging strategies. After many generations, a foraging animal takes actions to maximize the energy obtained per unit time spent foraging [31]. That is, poor foraging strategies are either eliminated or shaped into good ones. Escherichia coli (E. coli) is a common type of bacteria whose behavior to move comes from a set of up to six rigid 100–200 rps spinning flagella, each driven as a biological motor. An E. coli bacterium alternates between running and tumbling. The chemotactic actions of the bacteria are summarized as follows:

- If in a neutral medium, alternate tumbles and runs, its action is like performing a search.
- If swimming up a nutrient gradient (or out of noxious substances), or swim longer (climb up nutrient gradient or down noxious gradient) its behavior seeks increasingly favorable environments.
- If swimming down a nutrient gradient (or up noxious substance gradient), then search action is to avoid unfavorable environments.

2.6. DNA Computation

DNA computing is a form of computing which uses DNA (Deoxyribo-Nucleic Acid) and molecular biology, instead of the traditional silicon-based microprocessors. Just like a string of binary data is encoded with ones and zeros, a strand of DNA is encoded with four bases, represented by the letters A, T, C and G (nucleotides) and the data density is very impressive. An important property of DNA is its double stranded nature with every DNA sequence having a natural complement. This complementarity feature makes DNA a unique data structure for computation and can be exploited in many ways. In the cell, DNA is modified biochemically (on the molecular level) by a variety of enzymes, which are tiny protein machines. This mechanism along with some synthetic chemistry is responsible for the various DNA computation operators [33].

2.7. Computing with Words

Computing with words is a methodology in which the objects of computation are words and propositions drawn

from a natural language. Computing with words is inspired by the brain's crucial ability to manipulate perceptions without any measurements or computations. Computing with words provides a foundation for a computational theory of perceptions. A basic difference between perceptions and measurements is that, in general, measurements are crisp, whereas perceptions are fuzzy [34].

2.8. Simulated Annealing

Simulated annealing is based on the manner in which liquids freeze or metals re-crystalize in the process of annealing [20]. In an annealing process, molten metal, initially at high temperature, is slowly cooled so that the system at any time is approximately in thermodynamic equilibrium. If the initial temperature of the system is too low or cooling is done insufficiently slowly the system may become brittle or unstable with forming defects. The initial state of a thermodynamic system is set at energy Eand temperature T, holding T constant the initial configuration is perturbed and the change in energy dE is computed. If the change in energy is negative the new configuration is accepted. If the change in energy is positive it is accepted with a probability given by the Boltzmann factor exp -(dE/T). This processes is then repeated for few iterations to give good sampling statistics for the current temperature, and then the temperature is decremented and the entire process repeated until a frozen state is achieved at T=0.

2.9. Cultural Algorithms

Cultural Algorithms are computational models of cultural evolution. They consist of two basic components, a population space (using evolutionary algorithms), and a belief space. The two components interact by means of a vote-inherit-promote protocol [35]. Likewise the knowledge acquired by the problem solving activities of the population can be stored in the belief space in the form of production rules etc. Cultural algorithms represent a general framework for producing hybrid evolutionary systems that integrate evolutionary search and domain knowledge.

2.10. Artificial Life

Artificial life (alife) attempts at setting up systems with life like properties which all biological organisms possess, such as reproduction, homeostasis, adaptability etc. Alife is often described as attempting to understand high-level behavior from low-level rules; for example, how the simple rules of Darwinian evolution lead to high-level structure, or the way in which the simple interactions between ants and their environment lead to complex trailfollowing behavior [36]. Understanding this relationship in particular systems promises to provide novel solutions to complex real-world problems, such as disease prevention, stock-market prediction, and data-mining on the Internet.

2.11. Membrane Computing

Membrane computing (P systems) is a framework which abstracts from the way live cells process chemical compounds in their compartmental structure [21]. In a membrane system multisets of objects are placed in the compartments defined by the membrane structure, and the objects evolve by means of reaction rules are also associated within the compartments, and applied in a maximally parallel, nondeterministic manner. The objects can be described by symbols or by strings of symbols. For symbol-objects, a set of numbers are computed, and in the case of string-objects a set of strings are computed which is more like a language. The objects are able to through membranes and the membranes can dissolve, divide and change their permeability. These features are used in defining transitions between configurations of the system, and sequences of transitions are used to define computations. A sequence of transitions is a computation.

2.12. Quantum Computation

In conventional silicon computers, the amount of data is measured by bits; in a quantum computer, it is measured by qubits (quantum-bit). A qubit can be a 1 or a 0, or it can exist in a superposition that is simultaneously both 1 and 0 or somewhere in between. The basic principle of quantum computation is that the quantum properties of particles can be used to represent and structure data, and that devised quantum mechanisms can be used to perform operations with this data [22].

2.13. Hybrid Approaches

Several adaptive hybrid intelligent systems have in recent years been developed and many of these approaches use the combination of different knowledge representation schemes, decision making models and learning strategies to solve a computational task [23]. This integration aims at overcoming limitations of individual techniques through hybridization or fusion of various techniques. It is well known that the intelligent systems, which can provide human like expertise such as domain knowledge, uncertain reasoning, and adaptation to a noisy and time varying environment, are important in tackling practical computing problems. In contrast with conventional artificial intelligence techniques which only deal with precision and certainty the guiding principle of hybrid approaches is to exploit the tolerance for imprecision, uncertainty, robustness and to provide optimal solutions etc.



Figure 1. Web usage mining framework

3. Web Usage Mining and Business Intelligence

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the necessity of intelligent marketing strategies and relationship management.

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the web usage patterns could be directly applied to efficiently manage activities related to e-business, eservices and e-education. Accurate web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, tracking leaving customers etc. The usage information can be exploited to improve the performance of Web servers by developing proper prefetching and caching strategies so as to reduce the server response time. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content [12].

What makes the discovered knowledge interesting had been addressed by several works. Results previously known are very often considered as not interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpected appearance.

Web Usage Mining techniques can be used to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log. Recommendation systems could be developed to recommend interesting links to products which could be interesting to users.

One of the major issues in web log mining is to group all the users' page requests so to clearly identify the paths that users followed during navigation through the web site. The most common approach is to use cookies to track down the sequence of users' page requests or by using some heuristic methods. Session reconstruction is also difficult from proxy server log file data and sometimes not all users' navigation paths can be identified.



Figure 2. Web usage mining framework using natural computation [3]

Users' behavior can also be tracked down on the server side by means of TCP/IP packet sniffers.

Figure 1 [1] illustrates the basic Web usage mining framework. Whenever a visitor accesses the server it leaves, e.g., the IP, authenticated user ID, time/date, request mode, status, bytes, referrer and agent. The available data fields are specified by the HTTP protocol. There are several commercial software that could provide web usage statistics. These statistics could be useful for web administrators to get a sense of the actual load on the server. For small web servers, the usage statistics provided by conventional web site trackers may be adequate to analyze the usage patterns and trends. However, as the size and complexity of the data increases, the statistics provided by existing web log file analysis tools alone may prove inadequate and more knowledge mining and intelligent information processing will be necessary

3.1. Data Sources

The usage data collected at different sources represent the navigation patterns of different segments of the overall web traffic, ranging from single user, single site browsing behavior to multi-user, multi-site access patterns. Web server log does not accurately contain sufficient information for inferring the behavior at the client side as they relate to the pages served by the web server. Data may be collected from (a) Web servers, (b) proxy servers, and (c) Web clients. Web servers collect large amounts of information in their log files. Databases are used instead of simple log files to store information so to improve querying of massive log repositories [24]. Internet service providers use proxy server services to improve navigation speed through caching. Collecting navigation data at the proxy level is basically the same as collecting data at the server level but the proxy servers collects data of groups of users accessing groups of web servers. Usage data can be tracked also on the client side by using Javascript, Java applets, or even modified browsers [25].

3.2. Data Pre-Processing

The raw web log data after pre-processing and cleaning could be used for pattern discovery, pattern analysis, web usage statistics, and generating association/ sequential rules. Much work has been performed on extracting various pattern information from web logs and the application of the discovered knowledge range from improving the design and structure of a web site to enabling business organizations to function more efficiently.

Data pre-processing involves mundane tasks such as merging multiple server logs into a central location and parsing the log into data fields. The preprocessing comprises of (a) the data cleaning, (b) the identification and the reconstruction of users' sessions, and (c) the data formatting [26].

Data cleaning Consists of removing all the data tracked in Web logs that are useless for mining purposes. Graphic file requests, agent/spider crawling etc. could be easily removed by only looking for HTML file requests. Normalization of URL's is often required to make the requests consistent. For example requests for

www.okstate.edu
or
www.okstate.edu/index.html
are all for the same file.

Session identification and reconstruction consists of identifying the different users' sessions from the usually very poor information available in log files and reconstructing the users' navigation path within the identified sessions. Most of the problems encountered in this phase are caused by the caching performed either by proxy servers either by browsers. *Data formatting* is the final step of preprocessing in which data are properly formatted before applying data mining techniques.

Most of the Web usage mining applications research effort focuses on three main paradigms: association rules, sequential patterns, clustering and prediction [11]. Association rules are used to find associations among Web pages that frequently appear together in users' sessions. Sequential pattern discovery exposes sequential navigation patterns that appear in users' sessions frequently. Clustering has been widely used to group together similar sessions/access patterns. Prediction methods are used to forecast the volume of visitors / number of server accesses etc.

Jespersen et al. [9] proposed an hybrid approach for analyzing the visitor click sequences. A combination of Hypertext Probabilistic Grammar (HPG) and click fact table approach is used to mine web logs, which could be also used for general sequence mining tasks. In HPG, a non-terminal symbol corresponds to a web page and a production rule corresponds to a hypertext link.

Mobasher et al. [13] proposed the web personalization system that consists of offline tasks related to the mining of usage data and online process of automatic web page customization based on the knowledge discovered.

LOGSOM proposed by Smith et al. [17], utilizes selforganizing map to organize web pages into a twodimensional map based solely on the users' navigation behavior, rather than the content of the web pages.

LumberJack proposed by Chi et al. [6] builds up user profiles by combining both user session clustering and traditional statistical traffic analysis using K-means algorithm.

Joshi et al. [10] used relational online analytical processing approach for creating a web log warehouse using access logs and mined logs (association rules and clusters). Holden and Freitas [27] used an ant colony algorithm for Web page classification. A comprehensive overview of web usage mining research is found in [26][7].

In Figure 2, we present a web usage-mining framework using natural computing techniques [1]. Clustering of the log data helps to segregate similar visitors and the function approximation algorithm takes care of analyzing the visitor patterns and developing some trends. We started this research using a simple architecture consisting of a pre-processing block followed by self-organizing map (clustering) and neural network (function approximation). To improve the performance, we further investigated various hybrid combinations to cluster similar visitors based on the domain of origin, no of pages requested, time of access etc. For data clustering, self-organizing maps, evolutionary fuzzy clustering algorithm [2] and ant colony clustering algorithm [3] were used. The clustered data is then used to analyze the trends using several function approximation algorithms like neural networks; Takagi-Sugeno fuzzy inference system trained using neural network learning [4]; and linear genetic programming [5]. To demonstrate the efficiency of the proposed frameworks, web access log data at the Monash University's web site [14] were used for experimentations. Readers are advised to consult [1][2] and [3] for further details about the experiments and results.

4. Conclusions

This paper introduced the various ingredients and basic ideas of some of the important natural computation techniques. Further the importance of web usage mining in business intelligence was presented with an emphasis on data sources, data pre-processing and some related research.

Acknowledgements

This research was supported by the International Joint Research Grant of the IITA (Institute of Information Technology Assessment) foreign professor invitation program of the MIC (Ministry of Information and Communication), Korea.

References

- Abraham. A., Business Intelligence from Web Usage Mining, Journal of Information & Knowledge Management (JIKM), World Scientific Publishing Co., Singapore, Vol. 2, No. 4, pp. 375-390, 2003.
- [2] A. Abraham, "i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems", The IEEE International Conference on Fuzzy

Systems FUZZ-IEEE'03, USA, IEEE Press, pp. 1129-1134, 2003.

- [3] Abraham A. and Ramos V., Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming, 2003 IEEE Congress on Evolutionary Computation (CEC2003), Australia, IEEE Press, ISBN 0780378040, pp. 1384-1391, 2003.
- [4] A. Abraham, "Neuro-Fuzzy Systems: State-of-the-Art Modeling Techniques", in Jose Mira and Alberto Prieto, eds., Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence, Springer Verlag Germany, 2001, pp. 269-276.
- [5] W. Banzhaf, P. Nordin, E.R. Keller, and F.D. Francone, "Genetic Programming: An Introduction on The Automatic Evolution of Computer Programs and its Applications", Morgan Kaufmann Publishers, Inc., 1998
- [6] E.H. Chi, A. Rosien, and J. Heer, "LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition", In Proceedings of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM Press, 2002.
- [7] R. Cooley, "Web Usage Mining: Discovery and Application of Interesting patterns from Web Data", Ph. D. Thesis, University of Minnesota, Department of Computer Science, 2000.
- [8] J. Heer, and E.H. Chi, "Identification of Web User Traffic Composition using Multi- Modal Clustering and Information Scent", In Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51-58, 2001.
- [9] S.E. Jespersen, J. Thorhauge, and T. Bach, "A Hybrid Approach to Web Usage Mining, Data Warehousing and Knowledge Discovery", in Y. Kambayashi, W. Winiwarter, M. Arikawa, eds., LNCS 2454, pp. 73-82, 2002.
- [10] K.P. Joshi, A. Joshi, Y. Yesha, and R. Krishnapuram, "Warehousing and Mining Web Logs", Proceedings of the 2nd ACM CIKM Workshop on Web Information and Data Management, pp. 63-68, 1999.
- [11] R. Kosala, and H. Blockeel, Web Mining Research: A Survey, ACM SIGKDD Explorations, 2(1), pp. 1-15, 2000.

- [12] F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", Networking and Information Systems Journal (NIS), vol.2, no. 5-6, pp. 571-603, 1999.
- [13] Mobasher, R. Cooley, and J. Srivastava, "Creating Adaptive Web Sites through Usage-based Clustering of URLs", In Proceedings of 1999 Workshop on Knowledge and Data Engineering Exchange, USA, pp.19-25, 1999.
- [14] Monash University Web site: http://www.monash.edu.au
- [15] M. Perkowitz, and O. Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages", Proceedings of the 15th National Conference on Artificial Intelligence, pp. 727-732, 1998.
- [16] P. Pirolli, J. Pitkow, and R. Rao, "Silk From a Sow's Ear: Extracting Usable Structures from the Web", Proceedings on Human Factors in Computing Systems (CHI'96), ACM Press, 1996.
- [17] K.A. Smith and A. Ng, "Web page Clustering Using a Self-organizing Map of User Navigation Patterns", Decision Support Systems, Volume 35, Issue 2, pp. 245-256, 2003.
- [18] M. Spiliopoulou, and L.C. Faulstich, "WUM: A Web Utilization Miner", Proceedings of EDBT Workshop on the Web and Data Bases (WebDB'98), Springer Verlag, pp. 109-115, 1999.
- [19] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.
- [20] Kirkpatrick, S., C. D. Gelatt Jr., M. P. Vecchi, Optimization by Simulated Annealing, *Science*, 220, 4598, 671-680, 1983.
- [21] G. Paun, Computing with membranes, Journal of Computer and System Sciences, 61 (1), 108-143, 2000.
- [22] Deutsch, D., Quantum Theory, the Church-Turing Principle, and the Universal Quantum Computer". *Proc. Roy. Soc. Lond. A400*, 97–117, 1985.
- [23] A. Abraham, Intelligent Systems: Architectures and Perspectives, Recent Advances in Intelligent

Paradigms and Applications, Abraham A., Jain L. and Kacprzyk J. (Eds.), Studies in Fuzziness and Soft Computing, Springer Verlag Germany, ISBN 3790815381, Chapter 1, pp. 1-35, 2002.

- [24] K.P. Joshi, A. Joshi and Y. Yesha, On using a warehouse to analyze web logs, *Distributed and Parallel Databases*, **13** (2), pp. 161–180, 2003.
- [25] L.D. Catledge and J.E. Pitkow, Characterizing browsing strategies in the World-Wide Web, *Computer Networks and ISDN Systems* 27 (6), pp. 1065–1073, 1995.
- [26] F.M. Facca and P.L. Lanzi, Mining interesting knowledge from weblogs: a survey, Data & Knowledge Engineering, Volume 53, Issue 3, pp. 225-241, 2005.
- [27] N. Holden and A.A. Freitas. Web page classification with an ant colony algorithm. Proc. Parallel Problem Solving from Nature (PPSN-2004), LNCS 3242, pp. 1092-1102. Springer, 2004.
- [28] Bishop C.M., Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK, 1995.
- [29] Fogel, D. B., Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press, Piscataway, NJ, Second edition, 1999.
- [30] Kennedy J. and Eberhart R. Swarm intelligence. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2001.
- [31] Passino, K.M., Biomimicry of Bacterial Foraging for Distributed Optimization and Control, IEEE Control Systems Magazine, pp. 52-67, June 2002.
- [32] de Castro, L. N. and Timmis, J. I., Artificial Immune Systems: A New Computational Intelligence Approach, Springer-Verlag, London, 2002.
- [33] Amos M., Theoretical and Experimental DNA Computation. Springer, ISBN: 3-540-65773-8, 2005.
- [34] Zadeh L.A. and Kacprzyk J. (Eds.) Computing with Words in Information/Intelligent Systems: Foundations, Studies in Fuzziness and Soft Computing, Springer Verlag, Germany, ISBN 379081217X, 1999.
- [35] Reynolds R.G., Michalewicz, Z. Cavaretta M.J., Using Cultural Algorithms for Constraint Handling

in GENOCOP. Proceedings of the Fourth Annual Conference on Evolutionary Programming. MIT Press, Cambridge, pp. 289-305, 1995.

[36] C. Adami, Introduction to Artificial Life. Springer-Verlag New York, Inc., 1998.