Optimization of Location Allocation of Web Services Using A Multi-objective Particle Swarm Optimization

Boxiong Tan, Hui Ma, Mengjie Zhang

School of Engineering and Computer Science,
Victoria University of Wellington, New Zealand
{Boxiong.Tan, Hui.Ma, Mengjie.Zhang}@ecs.vuw.ac.nz

Abstract.

1 Introduction

There have been a growing number of interests in developing software systems unsing the service oriented archiecture (SOA). Service oriented archiecture is an archiecture strategy that enables Web application to be built using Web services. Web services are considered as self-describing, modular applications that can be published, located, and invoked across the Web [17]. Because of the convenience, low cost [1] and capacity to be composed into high-level business processes, Web service technology is becoming increasing popular.

Web services were hosted in heterogeneous server cluster or co-location centers that were widely distributed across different network providers and geographic regions. In recent years, web services are increasingly bing deployed in infrastructure-as-a-service (IaaS) clouds such as Amazon EC2, Windows Azure, and Rackspace, According to [8], 4% of Alexa‘s top 1 million use EC2/Azure and 96% web services were hosted in other platforms.

With the ever increasing number of functional similar web services being available on the Internet, the web service providers (WSPs) are trying to improve the quality of service (QoS) to become competitive in the market. QoS, also known as non-functional requirements to web services, is the degree to which a service meets specified requirements or user needs [23], such as response time, security and availability. Among numerous QoS measurements, service response time is a critical factor for many real-time services, e.g. traffic service or finance service. Service response time has two components: transmission time (variable with message size) and network latency [12]. Study [11] shows that network latency is a significant component of service response delay. Ignoring network latency will underestimate response time by more than 80 percent [19], since network latency is related to network topology as well as physical distance [10]. To reduce the network latency WSPs need to allocate services location where has the lower latency to the user center that access the services. Ideally, WSPs could deploy their services to each user center in order to provide the best quality. However, the more services deployed, the better the quality and the higher cost.

The Web service location-allocation problem is essentially a multi-objective optimization problem [2], for which there are two conflict objectives, to provide optimal QoS to service users and to consume minimal deployment cost. This problem is considered as NP-hard due to the fact that the combinatorial explosion of the search space [20].

Very few researches have studied the service location-allocation problem and most of the researchers treat this problem as a single objective problem. [1] [19] try to solve the problem by using integer linear programming techniques. In particular, [19] solved this problem by employing greedy and linear relaxations of Integer transpotation problem. However, the major problem for this approach is that linear programming is not scaling. Huang [9] proposed an enhanced genetic algorithm (GA)-based approach, which make use of the integer scalarization technique to solve this problem. This algorithm solves the problem with one objective and one constraint. However there are some deficiencies in the integer scalarization techniques [2]. Firstly, decision makers need to choose appropriate weights for the objectives to retrieve a satisfactorily solution. Secondly, non-convex parts of the Pareto set cannot be reached by minimizing convex combinations of the object functions.

Evolutionary multi-objective optimization (EMO) methodologies is ideal for solving multi-objective optimization problems [7], since EMO works with a population of solutions and a simple EMO can be extended to maintain a diverse set of solutions. With an emphasis for moving toward the true Pareto-optimal region, an EMO can be used to find multiple Pareto-optimal solutions in one single simulation run [13]. Among numerous EMO algorithms, Non-dominated sorting GA (NSGA-II) [5], Strength Pareto Evolutionary Algorithm 2 (SPEA-2) [6] have become standard approaches.

Several multiobjective optimization algorithms are based on Particle Swarm Optimization such as Multi-objective Particle Swarm Optimization (MOPSO) [3], Nondominated Sorting Particle Swarm Optimization (NSPSO) [15]. The performance of different multi-objective algorithms was compared in [3] using five test functions. These algorithms are NSGA-II, PAES [14], Micro-GA [4] and MOPSO. The results show that MOPSO was able to generate the best set of nondominated solutions close to the true Pareto front in all test functions except in one function where NSGA-II is superior.

Raquel and et al. [18] proposed a multi-objective Particle Swarm Optimization with crowding distance (MOPSOCD) that extends the MOPSO. The mechanism of crowding distance is incorporated into the algorithm specifically on global best selection and in the delection method of an external archive of nondominated solutions. The diversity of nondominated solutions in the external archive is amintained by using the mechanism of crowding distance together with a mutation operator. The performance shows that MOPSOCD is highly competitive in converging towards the Pareto front and has generated a well-distributed set of nondominated solutions.

In this paper, we proposed to use multi-objective to solve the web service location-allocation problem, which has two objectives, to minimize cost and de-

ployment network latency. We consider the problem faced by a WSP who has existing facilities but wishes to use the collected data to re-allocate their services in order to maximum their profit. The WSP must decide on facility locations from a finite set of possible locations. In order to make a decision, the WSP must first analyze the data collected from current use of services. The collected data should include the records of invocations from each unique IP address. Therefore, based on these data, the WSP could summarize several customer demands concentrated on n discrete nodes [1], namely user centers. We assume that the WSP has already done this step and a list of user centers and candidate service deployment locations are given. In addition to deciding locations to of the services, information about network latency between user centers and candidate locations is needed. Exist datasets in [21] [22] contain latency information collected from the real world.

The aim of this project is to propose a multi-objective PSO based approach to produce a set of near optimal solutions of service location-allocation, so that cost and overall network latency are close to minimum. Then, the service provider could use the algorithm which proposed by this paper, to select an optimal plan based on their funds. The main objectives are:

- To model the web service location-allocation problem so that it can be tackled by multi-objective PSO.
- To develop a multi-objective PSO based approach to the web service location-allocation problem.
- To evaluate our proposed approach using some existing datasets.

In Section 2 we introduce the background of multi-objective PSO and NSGA-II. In Section **??** we provide models of the service location allocation problems. Section **??** develops a MOPSOCD based algorithm. The experimental design and results evaluation are shown in Section **??**. Section **??** provides a brief summary.

2 Background

NSGA-II is a multi-objective algorithm based on genetic algorithm (GA) [16]. When used for problems with only two objectives, NSGA-II performs relatively well in both convergence and computing speed. It permits a remarkable level of flexibility with regard to performance assessment and design specification. NSGA-II assumes that every chromosome in the population has two attributes: a non-domination rank in the population and a local crowding distance in the population. The goal of NSGA-II is to converge to the Pareto front as possible and with even spread of the solutions on the front by controlling the two attributes.

The algorithm starts with a random initialization population. Once the population is sorted based on non-domination sorting, a rank is assigned to each chromosome. Then, a parameter called crowding distance is calculated for each individual. The crowding distance is a measure of how close an individual is to its neighbors. A large average crowding distance will result in better diversity in the population.

Parents are selected from the population by using tournament selection based on the rank and the crowding distance. An individual is selected in the rank if it is smaller than the other or if the crowding distance is greater than the other. The selected population generates offsprings using crossover and mutation operators.

The population with the current population and current offsprings is sorted again based on non-domination and only the best N individuals are selected, where N is the population size. The selection is based on rank and the on crowding distance on the last front.

References

1. Aboolian, R., Sun, Y., Koehler, G.J.: A locationallocation problem for a web services provider in a competitive market. European Journal of Operational Research 194(1), 64 – 77 (2009)
2. Caramia, M.: Multi-objective optimization. In: Multi-objective Management in Freight Logistics, pp. 11–36. Springer London (2008)
3. Coello, C., Pulido, G., Lechuga, M.: Handling multiple objectives with particle swarm optimization. Evolutionary Computation, IEEE Transactions on 8(3), 256–279 (June 2004)
4. Coello Coello Coello, C., Toscano Pulido, G.: A micro-genetic algorithm for multiobjective optimization. In: Zitzler, E., Thiele, L., Deb, K., Coello Coello, C., Corne, D. (eds.) Evolutionary Multi-Criterion Optimization, Lecture Notes in Computer Science, vol. 1993, pp. 126–140. Springer Berlin Heidelberg (2001)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on 6(2), 182–197 (2002)
6. Deb, K., Mohan, M., Mishra, S.: Evaluating the epsilon-domination based multiobjective evolutionary algorithm for a quick computation of pareto-optimal solutions. Evol. Comput. 13(4), 501–525 (2005)
7. Desai, S., Bahadure, S., Kazi, F., Singh, N.: Article: Multi-objective constrained optimization using discrete mechanics and nsga-ii approach. International Journal of Computer Applications 57(20), 14–20 (2012), full text available
8. He, K., Fisher, A., Wang, L., Gember, A., Akella, A., Ristenpart, T.: Next stop, the cloud: Understanding modern web service deployment in ec2 and azure. In: Proceedings of the 2013 Conference on Internet Measurement Conference. pp. 177–190. IMC '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2504730.2504740
9. Huang, H., Ma, H., Zhang, M.: An enhanced genetic algorithm for web service location-allocation. In: Decker, H., Lhotsk, L., Link, S., Spies, M., Wagner, R. (eds.) Database and Expert Systems Applications, Lecture Notes in Computer Science, vol. 8645, pp. 223–230. Springer International Publishing (2014)
10. Huffaker, B., Fomenkov, M., Plummer, D., Moore, D., claffy, k.: Distance Metrics in the Internet. In: IEEE International Telecommunications Symposium (ITS). pp. 200–202. IEEE, Brazil (Sep 2002)
11. Jamin, S., Jin, C., Kurc, A., Raz, D., Shavitt, Y.: Constrained mirror placement on the internet. In: INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. vol. 1, pp. 31–40 vol.1 (2001)

12. Johansson, J.M.: On the impact of network latency on distributed systems design. Inf. Technol. and Management 1(3), 183–194 (2000)
13. Kanagarajan, D., Karthikeyan, R., Palanikumar, K., Davim, J.: Optimization of electrical discharge machining characteristics of wc/co composites using non-dominated sorting genetic algorithm (nsga-ii). The International Journal of Advanced Manufacturing Technology 36(11-12), 1124–1132 (2008)
14. Knowles, J., Corne, D.: The pareto archived evolution strategy: a new baseline algorithm for pareto multiobjective optimisation. In: Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on. vol. 1, pp. –105 Vol. 1 (1999)
15. Li, X.: A non-dominated sorting particle swarm optimizer for multiobjective optimization. In: Cant-Paz, E., Foster, J., Deb, K., Davis, L., Roy, R., OReilly, U.M., Beyer, H.G., Standish, R., Kendall, G., Wilson, S., Harman, M., Wegener, J., Dasgupta, D., Potter, M., Schultz, A., Dowsland, K., Jonoska, N., Miller, J. (eds.) Genetic and Evolutionary Computation GECCO 2003, Lecture Notes in Computer Science, vol. 2723, pp. 37–48. Springer Berlin Heidelberg (2003), http://dx.doi.org/10.1007/3-540-45105-6_4
16. Man, K.F., Tang, K.S., Kwong, S.: Genetic algorithms: concepts and applications. IEEE Transactions on Industrial Electronics 43(5), 519–534 (1996)
17. Ran, S.: A model for web services discovery with QoS. SIGecom Exch. 4(1), 1–10 (2003)
18. Raquel, C.R., Naval, Jr., P.C.: An effective use of crowding distance in multiobjective particle swarm optimization. In: Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation. pp. 257–264. GECCO '05, ACM, New York, NY, USA (2005)
19. Sun, Y., Koehler, G.J.: A location model for a web service intermediary. Decis. Support Syst. 42(1), 221–236 (2006)
20. Vanrompay, Y., Rigole, P., Berbers, Y.: Genetic algorithm-based optimization of service composition and deployment. In: Proceedings of the 3rd International Workshop on Services Integration in Pervasive Environments. pp. 13–18. SIPE '08, ACM (2008)
21. Zhang, Y., Zheng, Z., Lyu, M.: Exploring latent features for memory-based QoS prediction in cloud computing. In: Reliable Distributed Systems (SRDS), 2011 30th IEEE Symposium on. pp. 1–10 (2011)
22. Zheng, Z., Zhang, Y., Lyu, M.: Distributed QoS evaluation for real-world web services. In: Web Services (ICWS), 2010 IEEE International Conference on. pp. 83–90 (2010)
23. Zhou, J., Niemela, E.: Toward semantic QoS aware web services: Issues, related studies and experience. In: Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on. pp. 553–557 (2006)